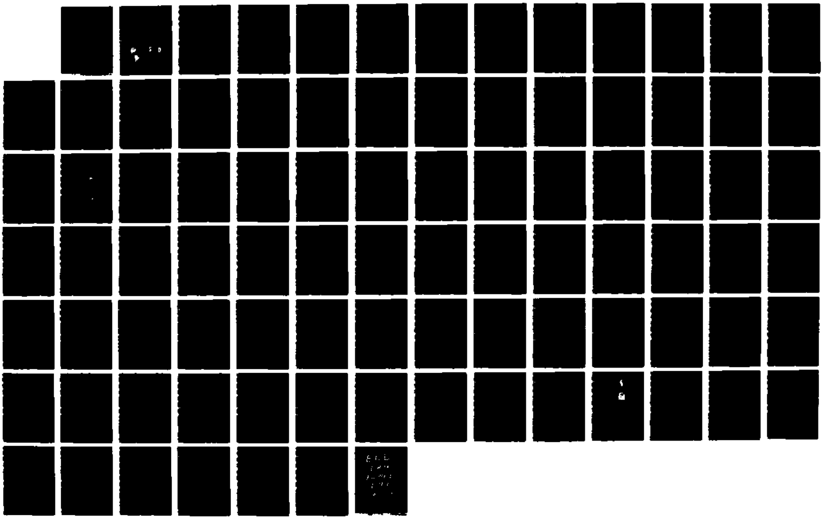
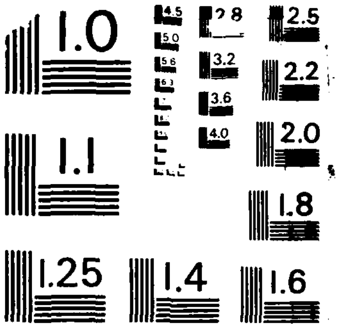
AD-A190 435 ADVANCED GUIDANCE ALGORITHMS FOR HOMING MISSILES WITH BEARINGS-ONLY MEASU (U) TEXAS UNIV AT AUSTIN DEPT OF AEROSPACE ENGINEERING AND ENGINE J L SPEYER ET AL 1/1

UNCLASSIFIED NOV 87 AFOSR-RR-87-1962 SAFOSR-84-0371 F/G 17/7.3 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

2

THE UNIVERSITY OF TEXAS AT AUSTIN

Final Technical Report

Grant AFOSR-84-0371

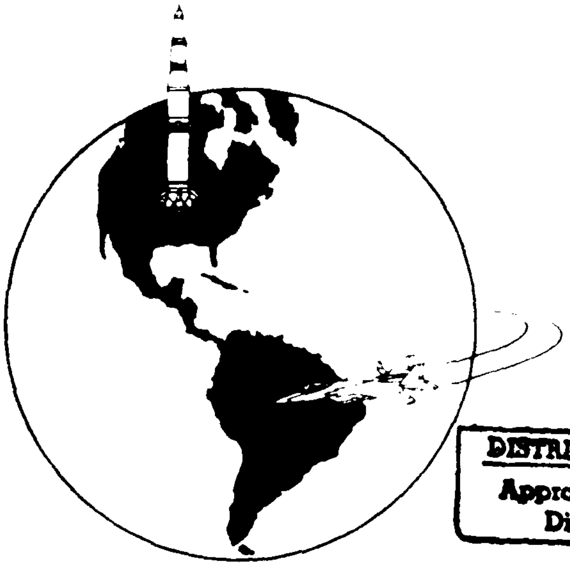
Advanced Guidance Algorithms for Homing Missiles With Bearings-Only Measurements

by

Jason L. Speyer and David G. Hull

Guidance and Control Group

November, 1987



Department of Aerospace Engineering and Engineering Mechanics

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TK- 87-1962

6a. NAME OF PERFORMING ORGANIZATION
University of Texas

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
AFOSR/NM

6c. ADDRESS *(City, State, and ZIP Code)*
Austin, TX

7b. ADDRESS *(City, State, and ZIP Code)*
AFOSR/NM
Bldg 410
Bolling AFB DC 20332-6448

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL *(If applicable)*
NM

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-84-0371

8c. ADDRESS *(City, State, and ZIP Code)*
AFOSR/NM
Bldg 410
Bolling AFB DC 20332-6448

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO.
61102F	2304	A1	

11. TITLE *(Include Security Classification)*
Advanced Guidance Algorithms for Homing Missiles with Bearings-Only Measurements

12. PERSONAL AUTHOR(S)
Jason L. Speyer & David G. Hull

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM ________ TO ______

14. DATE OF REPORT *(Year, Month, Day)*
Nov. 1987

15. PAGE COUNT

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Homing missiles guidance is formulated as an optimal stochastic control problem where the special nonlinear structure of the missile-target engagement is exploited. Since this stochastic control problem assumes a nested information pattern, the filter structure can be developed independently of the guidance scheme. However, the guidance schemem is dependent on and affects filter performance. Significant progress is being made on both the estimatio problem and the guidance problem. Investigation of the nonlinear estimators especially tailored to the homing missile problem has produced not only a good deal of insight but responsive and mechanizable schemes. Although these schemes are applicable to active sensor our emphasis has been on the more difficult passive sensor case where only angles are available. Recently-developed schemes based on coordinate transformations and on an assumed probability density function perform well, but the modified-gain extended Kalman filter seems to be the most promising.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
Maj. James Crowley

22b. TELEPHONE *(Include Area Code)*
(202) 767-5025

22c. OFFICE SYMBOL
NM

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Final Technical Report
Grant AFOSR-84-0371

Advanced Guidance Algorithms for Homing Missiles With Bearings-Only Measurements

by

Jason L. Speyer and David G. Hull
Guidance and Control Group
Department of Aerospace Engineering and Engineering Mechanics
University of Texas at Austin

November, 1987

Summary

Homing missile guidance is formulated as an optimal stochastic control problem where the special nonlinear structure of the missile-target engagement is exploited. Since this stochastic control problem assumes a nested information pattern, the filter structure can be developed independently of the guidance scheme. However, the guidance scheme is dependent on and affects filter performance. Significant progress is being made on both the estimation problem and the guidance problem.

Investigation of the nonlinear estimators especially tailored to the homing missile problem has produced not only a good deal of insight but responsive and mechanizable schemes. Although these schemes are applicable to active sensors, our emphasis has been on the more difficult passive sensor case where only angles are available. Recently-developed schemes based on coordinate transformations and on an assumed probability density function perform well, but the modified-gain extended Kalman filter seems to be the most promising. Furthermore, this filter has been used as the basis of a stochastic adaptive flight control scheme. In order to improve this class of stochastic control schemes, new results have been obtained in control synthesis for structured plant uncertainties.

Two important current efforts in missile guidance with bearings-only information are in the development of the guidance schemes that enhance an information measure by trajectory modulation and in target acceleration detection. A mechanizable guidance law based upon linear-quadratic-Gaussian theory which modulates the path initially to enhance the information measure but which meets terminal miss constraints has been tested. Finally, based upon deterministic detection filter design by spectral methods, a detection scheme for rapidly detecting target motion has been developed and is being compared with current designs.

Research Objectives and Status

A special class of stochastic control systems is being developed for the guidance system of a homing missile by exploiting the special nonlinear structure of the missile-target engagement. Improvements are required in the current guidance law, proportional navigation, because the guidance system degrades under initial intercept geometries that produce large nonlinearities about the homing triangle or due to active target motion which also induces large nonlinearities. Our guidance law investigations have emphasized measurements from passive sensors for which only bearing information is available. This bearings-only guidance problem is most challenging because the stochastic controller has the dual role of enhancing filtler performance and achieving target intercept with minimal expected terminal cost. However, this problem is somewhat simplified since the separation theorem in the sense of Witsenhausen is satisfied. The separation theorem states that the filter structure, given the classical information pattern, is independent of the controller structure although the controller is highly dependent on the predicted filter performance.

Motivated by the separation theorem, high-performance estimators have been developed which are tailored to the special nonlinearities of the missile-target engagement. One new estimator, called the modified-gain extended Kalman filter (MGEKF), is applicable to two important engineering problems: bearings-only estimation [1] and state and parameter estimation [2]. Although we consider the MGEKF a breakthrough in guidance filter development, the assumed-density filter [3] and the coordinate-transformation-based filter [4] have also shown considerable promise. Since the conditional mean estimator is infinite dimensional, the finite-dimensional MGEKF is proposed as the estimation processor for the homing guidance dual controller. Furthermore, the MGEKF is also proposed as the state and parameter estimator for an explicit adaptive control law which is applicable to flight control and autopilot design. In particular, the MGEKF has been applied to the problem of on-line state estimation and the identification of aircraft stability derivatives [5]. An adaptive control loop using this estimator is given in [6] where the essential parameter required is moment coefficient due to elevator defection. The adaptive gain is inveresely proportional to this parameter which seems well estimated by the MGEKF even in moderately-high clear air turbulence. However, more elaborate controllers will be required for bank-to-turn missiles. A multivariable synthesis scheme is suggested in [7] in which the LQG controller can be made insensitive to a class of parameter variations. It is seen in [5,6] that the moment coefficients are estimated well but the force coefficents are not. In particular, their estimation response is quite sluggish due to the effect of high-frequency noise associated with the model of the clear air turbulence. An adaptive system is being designed so that the controller is only sensitive to the moment coefficients. This approach to autopilot design is being considered for application to a bank-to-turn missile.

Both homing missile guidance and adaptive control schemes are currently designed based upon the certainty equivalence principle. That is, a controller and estimator are placed in cascade where both are designed independently of one another. These ad hoc controller structures are not adequate in general, and improvements are sought through the dual control concept. The dual controller structure which has never been realized by even the simplest stochastic control example needs much study. We began our efforts by noting that the essence of the dual control problem is captured in deterministic setting where the nonlinear observer performance is enhanced by trajectory modulation. In particular, a measure associated with the Fisher information matrix is maximized in order to obtain an information-enhanced homing path [8,9,10]. In [10] not only is the EKF performance improved by trajectory modulation over the proportional navigation path, but the performance of the MGEKF along these information-enhanced paths relative to that of the EKF is impressive.

Based on these results an ad hoc guidance rule which seems to possess the dual control property is proposed. It is seen that the trace of the information matrix weighted by the range-to-go when combined with the current control performance index reduces to a quadratic form. This form differs from current forms in that the performance index due to the information measure is *not* convex. Some preliminary results are given in [11]. It is noted that this simple guidance rule produces trajectories similar to those generated in [8,9,10].

The essential difficulty in dealing with dual control problems is that the sturcture of the controller is not well understood. For this reason, ad hoc schemes pervade the literature, but no rational scheme is ever suggested. For this reason, we have begun looking into asymptotic approaches to this class of problems. For small measurement and process noise variances, the optimal control law, obtained from the Hamilton-Jacobi-Bellman PDE of a particular nonlinear problem, is determined in terms of an asymptotic expansion in the state estimate and state error variance. This problem is chcsen because the estimation process is conditionally Gaussian and the deterministic problem (or zeroth-order solution of the Hamilton-Jacobi-Bellman equation) is integrable. Since it is hypothesized that dual control problems are not integrable, the expansion about the zeroth-order solution should give valuable insight into the structure of the dual control problem. The objective is to apply these ideas to both the homing guidance and adaptive autopilot problems.

There is a real need to determine the effects of guidance system errors on missile guidance. To do this, a measure of performance is used in [12] which is associated with the optimal return function of the LQG problem and has the property of a Lyapunov function. Since the guidance laws considered to date are based upon the certainty equivalence principle, the control is a function of the filter or observer output. The Lyapunov function is a function of three terms, one associated with the LQ problem, one associated with the observer, and one associated with the error in the control law due to the inaccuracy of the state estimate from the observer.

Finally, the very important problem of target maneuver detection is considered. Our approach is to develop target motion sensitive filters (actually observers). The theory has been developed for time invariant linear dynamic systems [13,14,15]. The objective is to design the detection gain so that the target motion can be associated directly with the measurement residuals. Our present effort is described in [15].

References

1. T. L. Song and J. L. Speyer, "A Stochastic Analysis of a Modified Gain Extended Kalman Filter with Applications to Estimation with Bearings Only Measurements," *IEEE Trans. on Auto. Control*, Vol. AC-30, No. 10, October, 1985, pp. 940-949.

2. T. L. Song and J. L. Speyer, "The Modified Gain Extended Kalman Filter and Parameter Identification in Linear Systems," *Automatica*, Vol. 22, No. 1, January, 1986, pp. 59-75.

3. S. N. Balakrishman and J. L. Speyer, "Assumed Density Filter with Application to Homing Missile Guidance," *Proceedings of the AIAA Guidance and Control Conference*, Williamsburg, Virginia, August, 1986, pp. 905-914; and to be published in the AIAA Journal of Guidance, Control and Dynamics.

4. S. N. Balakrishnan and J. L. Speyer, "Coordinate-Transformation-Based Filter for Improved Target Tracking," *AIAA Journal of Guidance, Control, and Dynamics*, Vol. 9, No. 6, November-December, 1986, pp. 704-709.

5. J. L. Speyer and E. Z. Crues, "On-Line Aircraft State and Stability Derivation Estimation Using the Modified Gain Extended Kalman Filter,", *AIAA Journal of Guidance, Control, and Dynamics*, Vol. 10, No. 3, May-June, 1987, pp. 262-268.

6. J. L. Speyer, E. Z. Crues and W. P. Fun, "An Explicit Adaptive Flight Control System Based on the Modified Gain Extended Kalman Filter," *Proceedings of the AIAA Guidance and Control Conference*, Williamsburg, Virginia, August, 1986, pp. 573-580.

7. M. Tahk and J. L. Speyer, "Modeling of Parameter Variations and Asymptotic LQG Synthesis," *IEEE Trans. on Auto. Control*, Vol. AC-32, No. 9, September, 1987, pp. 793-801.

8. J. L. Speyer, D. G. Hull, C. Y. Tseng and S. W. Larson, "Estimation Enhancement by Trajectory Modulation for Homing Missiles,"*AIAA*

Journal of Guidance, Control, and Dynamics, Vol. 7, No. 2, March-April, 1984, pp. 167-174.

9. D. G. Hull, J. L. Speyer and C. Y. Tseng, "Maximum Information Guidance for Homing Missiles," *AIAA Journal of Guidance, Control and Dynamics*, Vol. 8, No. 4, July-August, 1985, pp. 494-497.

10. J. L. Speyer, D. G. Hull and W. P. Bernard, "Performance of the Modified-Gain Extended Kalman Filter Along an Enhanced Information Path of a Homing Missile," *Proceedings of the AIAA Guidance and Control Conference*, Williamsburg, Virginia, August, 1986, pp. 897-904.

11. D. G. Hull, J. L. Speyer and D. B. Burris, "A Linear-Quadratic Guidance Law for the Dual Control of Homing Missiles," *Proceedings of the AIAA Guidance and Control Conference*, Monterey, California, August, 1987, pp. 551-559.

12. P. L. Vergez, "Closed- Loop System Analysis using Lyanpunov Stability Theory," Ph. D. Dissertation, The University of Texas at Austin, May, 1987.

13. J. E. White and J. L. Speyer, "Detection Filter Design: Spectral Theory and Algorithms," *IEEE Trans. on Auto. Cont.*, Vol. AC-32, No. 7, July, 1987, pp. 593-603.

14. S. Attar, "Analysis and Design of Detection Filters with Closed-Loop Systems," Ph. D. Dissertation, The University of Texas at Austin, August, 1987.

15. G. A. Bowman and J. L. Speyer, "Detection Filters for Missile Tracking," *Proceedings of the AIAA Guidance and Control Conference*, Monterey, California, 1987, pp. 570-578.

Professional Personnel Associated with the Research Effort

M. S. Students

1. W. P. Bernard
2. G. A. Bowman
3. D. B. Burris
4. E. Z. Crues
5. W. P. Fun
6. C. R. Jaensch
7. R. A. Luke

Ph. D. Students

1. S. Attar
2. Y. Hahn
3. D. Kim
4. P. L. Vergez
5. J. E. White

Faculty

1. D. G. Hull
2. J. L. Speyer

Advanced Degrees Awarded

M. S. Degrees

1. Luke, R. A., "Improvement of Target State Estimation for an Air-to-Air Missile Via Kalman Filtering," December, 1984.

2. Crues, E. Z., "Use of the Modified Gain Extended Kalman Filter in the Identification of Aircraft Stability Derivatives," May, 1985.

3. Bernard, W. P. H., "Filter Enhancement for Air-to Air Missiles with Angle-Only Measurements," December, 1985.

4. Fun, W. P., "An Explicit Adaptive Flight Control System Based on the Modified Gain Extended Kalman Filter," December, 1986.

5. Bowman, G. A., "Detection Filters for Missile Tracking," May, 1987.

Ph. D. Degrees

1. White, J. E., "Detection Filter Design by Eigensystem Assignment," May, 1985.

2. Tahk, M. J., "Design Synthesis Techniques for Multivariable Linear Systems under Structured Parameter Variations," December, 1986.

3. Vergez, P. L., "Closed-Loop System Analysis using Lyapunov Stability Theory," May, 1987.

4. Attar, S., "Analysis and Design of Detection Filters Within Closed-Loop Systems," August, 1987.

Chronological List of Journal Publications

A chronological list of journal articles published during the grant period is the following:

1. D. G. Hull, J. L. Speyer and C. Y. Tseng, "Maximum Information Guidance for Homing Missiles," *AIAA Journal of Guidance, Control and Dynamics*, Vol. 8, No. 4, July-August, 1985, pp. 494-497.

2. T. L. Song and J. L. Speyer, "A Stochastic Analysis of a Modified Gain Extended Kalman Filter with Applications to Estimation with Bearings Only Measurements," *IEEE Trans. on Auto. Control*, Vol. AC-30, No. 10, October, 1985, pp. 940-949.

3. T. L. Song and J. L. Speyer, "The Modified Gain Extended Kalman Filter and Parameter Identification in Linear Systems," *Automatica*, Vol. 22, No. 1, January, 1986, pp. 59-75.

4. S. N. Balakrishman and J. L. Speyer, "Assumed Density Filter with Application to Homing Missile Guidance," *Proceedings of the AIAA Guidance and Control Conference*, Williamsburg, Virginia, August, 1986, pp. 905-914; and to be published in the AIAA Journal of Guidance, Control and Dynamics.

5. S. N. Balakrishnan and J. L. Speyer, "Coordinate-Transformation-Based Filter for Improved Target Tracking," *AIAA Journal of Guidance, Control, and Dynamics*, Vol. 9, No. 6, November-December, 1986, pp. 704-709.

6. J. L. Speyer and E. Z. Crues, "On-Line Aircraft State and Stability Derivation Estimation Using the Modified Gain Extended Kalman Filter,", *AIAA Journal of Guidance, Control, and Dynamics*, Vol. 10, No. 3, May-June, 1987, pp. 262-268.

7. J. E. White and J. L. Speyer, "Detection Filter Design: Spectral Theory and Algorithms," *IEEE Trans. on Auto. Cont.*, Vol. AC-32, No. 7, July, 1987, pp. 593-603.

8. M. Tahk and J. L. Speyer, "Modeling of Parameter Variations and Asymptotic LQG Synthesis," *IEEE Trans. on Auto. Control*, Vol. AC-32, No. 9, September, 1987, pp. 793-801.

Copies of these articles are included in the Appendix.

Appendix

Maximum-Information Guidance for Homing Missiles

D.G. Hull,* and J.L. Speyer*
University of Texas at Austin, Austin, Texas
and
C.Y. Tseng†
Chung Shan Institute of Science and Technology, Taipei, Taiwan, Republic of China

A recently-defined information index is used to enhance the information content of minimum-control-effort trajectories for the homing missile intercept problem. Optimal planar intercept trajectories are obtained for a performance index which is control effort weighted by position information content. The missile and target are assumed to be operating at constant speed. The shooting method is used to compute the optimal paths; but because of the simplicity of the model, on-line optimization yielding a guidance law with information enhancement should be possible.

Nomenclature

A	$=\cos\phi/v_R$
a	= missile normal acceleration (ft/s^2)
B	$=\sin\phi/v_R$
c	= constant in measurement variance model (ft^{-2})
G	= augmented end-point function
H	= variational Hamiltonian
R	= range (ft)
t	= time (s)
v_R	= ratio of missile velocity to target velocity
V	= velocity
W	= weight
X,Y	= planar coordinates (ft)
α	= nondimensional missile normal acceleration
θ	= missile velocity angle
λ	= time-dependent Lagrange multiplier
ν	= constant Lagrange multiplier
ξ,η	= nondimensional relative coordinates
ρ	= nondimensional range
τ	= nondimensional time
ϕ	= missile velocity angle

Superscripts

$(\dot{\ })$	= derivative with respect to t
$(\ )'$	= derivative with respect to τ

Subscripts

f	= final point
M	= missile
R	= relative
T	= target
0	= initial point

Introduction

IN Ref. 1, the problem of enhancing the information content of angle measurements in a homing missile engagement is considered. While the dynamics used in the filter development are linear in the states (relative position, relative velocity, and target acceleration), the measurements are nonlinear in a rectangular coordinate frame. Hence, the trajectory followed by the missile affects the measurement sequence and, in turn, the ability of the filter to extract the states from the measurements. A scalar performance index representing a measure of the information content of the missile path is developed, and a maximum-information intercept trajectory is determined. Next, measurements are created along the maximum-information path and processed with an extended Kalman filter. It is shown that the filter performs considerably better for measurements made along the maximum-information path than it does for measurements made along a proportional-navigation path. In fact, the filter diverges from the true state along the proportional-navigation path and converges to the true state along the maximum-information path.

Since the trajectory determined from the scalar information performance index reported in Ref. 1 induces greatly improved state estimation results, its use in the development of an information-enhancement guidance law is investigated. Because of the complexity of the problem, the simplest-possible intercept problem is formulated, that is, two-dimensional motion of a constant velocity missile and target. The performance index is taken to be the control effort weighted by the information index, and solutions are obtained by the shooting method. However, to obtain initial values of the Lagrange multipliers required by the shooting method, the problem of minimizing just the control effort must be considered first. Then, by solving the weighted problem in stages (gradually increasing the weight from zero), the desired optimal trajectories can be obtained.

Statement of the Problem

The classical guidance law known as proportional navigation is a perturbation guidance law about a nominal intercept triangle. The intercept triangle is essentially a minimum-control-effort trajectory in the plane (see Fig. 1 for nomenclature). For a constant velocity, steerable missile and a constant velocity target moving in a straight line, the optimal control problem is stated as follows:

Find the missile normal-acceleration history $a(t)$ which minimizes

$$J=\frac{1}{2}\int_{t_0}^{t_f} a^2\mathrm{d}t \tag{1}$$

Received June 11, 1984; presented as Paper 84-1887 at the AIAA Guidance and Control Conference, Seattle, Wash., Aug. 20-22, 1984; revision received Sept. 20, 1984.

*Professor, Department of Aerospace Engineering and Engineering Mechanics. Associate Fellow AIAA.

†Research Associate.

subject to dynamical constraints

$$\dot{X}_R = V_T\cos\phi - V_M\cos\theta$$

$$\dot{Y}_R = V_T\sin\phi - V_M\sin\theta$$

$$\dot{\theta} = a/V_M \tag{2}$$

and the prescribed boundary conditions

$$t_0 = 0, \quad X_{R_0} = R_0, \quad Y_{R_0} = 0, \quad \theta_0 \equiv \text{free}$$

$$t_f \equiv \text{free}, \quad X_{R_f} = 0, \quad Y_{R_f} = 0, \quad \theta_f \equiv \text{free} \tag{3}$$

This optimal control problem[2] admits the solution $a = 0$ or $\theta = \text{const}$. However, along this path, the filter is not able to estimate all of the states because the range along the line-of-sight is unobservable. To enhance state estimation, it is possible to weight the final time with a term associated with information content. The simplest form of this term is obtained by considering only the position-information part of the performance index developed in Ref. 1. With this term included, the performance index, Eq. (1), can be rewritten as

$$J = \frac{1-W}{2}\int_{t_0}^{t_f} a^2 \mathrm{d}t - W\int_{t_0}^{t_f} \frac{\mathrm{d}t}{1 + c(X_R^2 + Y_R^2)} \tag{4}$$

where W is the weight and c is a constant associated with the measurement variance model used in the filter. If $W = 0$, J is the control effort; and if $W = 1$, it becomes the information integral. Since a minimum is being sought and since the information is to be maximized, the minus sign is introduced to convert the maximization problem to a minimization problem. Finally, when actually implemented, it is envisioned that W would be related to the state estimation error covariance, increasing as the covariance increases.

At this point, the following nondimensional variables are introduced:

$$\xi = \sqrt{c}X_R, \quad \eta = \sqrt{c}Y_R, \quad \tau = \sqrt{c}V_M t,$$

$$v_R = V_M/V_T, \quad \alpha = a/cV_M^2, \quad \rho = \sqrt{c}R \tag{5}$$

In terms of these variables, the optimal control problem is to find the missile normal-acceleration history $\alpha(\tau)$ which minimizes the performance index

$$J = \frac{1-W}{2}\int_{\tau_0}^{\tau_f} \alpha^2 \mathrm{d}\tau - W\int_{\tau_0}^{\tau_f} \frac{\mathrm{d}\tau}{1 + \xi^2 + \eta^2} \tag{6}$$

subject to the system dynamics

$$\xi' = \cos\phi/v_R - \cos\theta,$$

$$\eta' = \sin\phi/v_R - \sin\theta,$$

$$\theta' = \alpha \tag{7}$$

and the prescribed boundary conditions

$$\tau_0 = 0, \quad \xi_0 = \rho_0, \quad \eta_0 = 0, \quad \theta_0 \equiv \text{free} \tag{8a}$$

$$\tau_f \equiv \text{free}, \quad \xi_f = 0, \quad \eta_f = 0, \quad \theta_f \equiv \text{free} \tag{8b}$$

This optimal control problem does not yield an analytical solution and is solved with the numerical optimization method known as the shooting method. Because of the sensitivity of the shooting method to initial guesses, the problem is solved analytically for $W = 0$ to obtain Lagrange multipliers. Then, with these multipliers as initial guesses, the shooting method is converged for a small value of W. W is increased, and the process is repeated with the last converged multipliers as initial guesses.

Minimum Control-Effort Problem

For the case where $W = 0$, the variational Hamiltonian and the augmented end-point function are given by

$$H = \alpha^2/2 + \lambda_1(A - \cos\theta) + \lambda_2(B - \sin\theta) + \lambda_3\alpha$$

$$G = \nu_1\xi_f + \nu_2\eta_f \tag{9}$$

where $\lambda_i (i = 1,2,3)$ is a time-varying Lagrange multiplier, $\nu_i (i = 1,2)$ is a constant Lagrange multiplier, $A = \cos\phi/v_R$, and $B = \sin\phi/v_R$. The Euler-Lagrange equations[2] for λ lead to

$$\lambda_1' = -H_\xi = 0 \tag{10a}$$

$$\lambda_2' = -H_\eta = 0 \tag{10b}$$

$$\lambda_3' = -H_\theta = -\lambda_1\sin\theta + \lambda_2\cos\theta \tag{10c}$$

where the optimal control satisfies the optimality condition

$$H_\alpha = \alpha + \lambda_3 = 0 \tag{11}$$

Finally, the natural boundary conditions are

$$\lambda_{1_f} = G_{\xi_f} = \nu_1$$

$$\lambda_{2_f} = G_{\eta_f} = \nu_2$$

$$\lambda_{3_f} = G_{\theta_f} = 0, \quad \lambda_{3_0} = G_{\theta_0} = 0$$

$$H_f = \alpha_f^2/2 + \lambda_{1_f}(A - \cos\theta_f)$$
$$+ \lambda_{2_f}(B - \sin\theta_f) + \lambda_{3_f}\alpha_f = 0 \tag{12}$$

It is observed that the absolute minimum control effort is achieved when $\alpha = 0$. Whether or not this can be the solution is now investigated. Equations (10a) and (10b) indicate that λ_1 and λ_2 are constants so that Eq. (10c) gives $\theta = \text{const}$. Hence, the system equations, Eq. (7), can be integrated subject to the final conditions of Eq. (8b) to obtain

$$0 = (A - \cos\theta)\tau_f + \xi_0, \qquad 0 = B - \sin\theta \tag{13}$$

which determines θ and τ_f as follows:

$$\sin\theta = B, \qquad \tau_f = \xi_0/(\cos\theta - A) \tag{14}$$

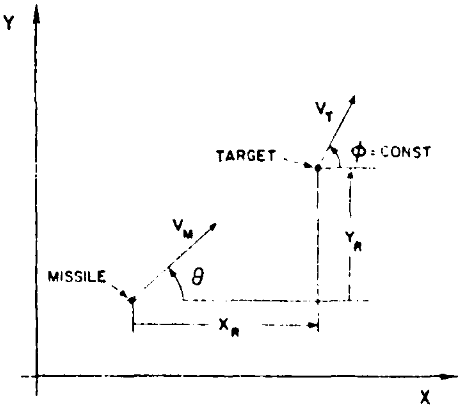


Fig. 1 Two-dimensional intercept geometry.

Next, Eq. (11) gives $\lambda_3 = 0$ which satisfies the natural boundary conditions of Eq. (12). Finally, Eqs. (10) and (12) lead to

$$\lambda_1 = 0, \qquad \lambda_2 = 0 \tag{15}$$

These values of λ will be used to begin the solution of the information-weighted minimum control-effort problem.

Minimum Information-Weighted Control-Effort Problem

For $W \neq 0$, the variational Hamiltonian and the augmented endpoint functions are defined as

$$H = \frac{1-W}{2}\alpha^2 - \frac{W}{1+\xi^2+\eta^2} + \lambda_1(A-\cos\theta) + \lambda_2(B-\sin\theta) + \lambda_3\alpha$$

$$G = \nu_1\xi_f + \nu_2\eta_f$$

Next, the differential equations for the λ's are given by

$$\lambda_1' = -2W\xi/(1+\xi^2+\eta^2)^2$$

$$\lambda_2' = -2W\eta/(1+\xi^2+\eta^2)^2$$

$$\lambda_3' = -\lambda_1\sin\theta + \lambda_2\cos\theta \tag{17}$$

while the optimal control must satisfy

$$(1-W)\alpha + \lambda_3 = 0 \tag{18}$$

Finally, the natural boundary conditions lead to

$$\lambda_{1_f} = \nu_1, \quad \lambda_{2_f} = \nu_2, \quad \lambda_{3_f} = 0, \quad \lambda_{3_0} = 0$$

$$-\frac{W}{1+\xi_f^2+\eta_f^2} + \lambda_{1_f}(A-\cos\theta_f) + \lambda_{2_f}(B-\sin\theta) = 0 \tag{19}$$

Unfortunately, this optimal control problem does not yield an analytical result so that numerical methods must be employed. Here, the shooting method[4] is used to solve the corresponding two-point boundary-value problem (TPBVP). It is formed by solving Eq. (18) for the control and eliminating α from the remaining equations to obtain the differential system

$$\xi' = A - \cos\theta$$

$$\eta' = B - \sin\theta$$

$$\theta' = -\lambda_3/(1-W)$$

$$\lambda_1' = -2W\xi/(1+\xi^2+\eta^2)^2$$

$$\lambda_2' = -2W\eta/(1+\xi^2+\eta^2)^2$$

$$\lambda_3' = -\lambda_1\cos\theta + \lambda_2\sin\theta \tag{20}$$

and the boundary conditions

$$\tau_0 = 0, \qquad \xi_0 = \rho_0, \qquad \eta_0 = 0, \qquad \lambda_{3_0} = 0$$

$$\xi_f = 0, \qquad \eta_f = 0, \qquad \lambda_{3_f} = 0$$

$$-\frac{W}{1+\xi_f^2+\eta_f^2} + \lambda_{1_f}(A-\cos\theta_f) + \lambda_{2_f}(B-\sin\theta_f) = 0 \tag{21}$$

The TPBVP is solved by using the initial Lagrange multipliers for $W = 0$ and a small value of W. Then, as W is increased, the initial guess for the λ's is the converged set for the

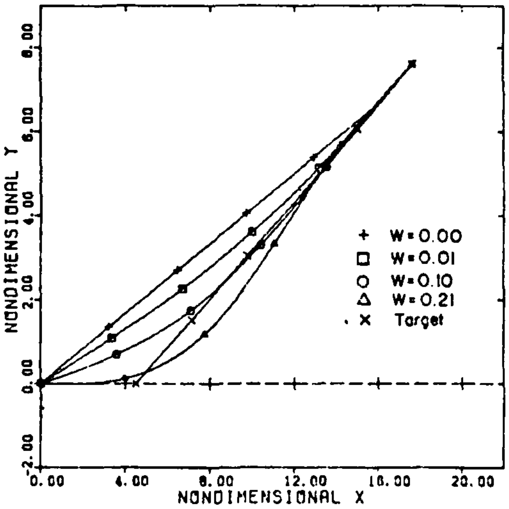


Fig. 2 Information-enhanced optimal intercept paths.

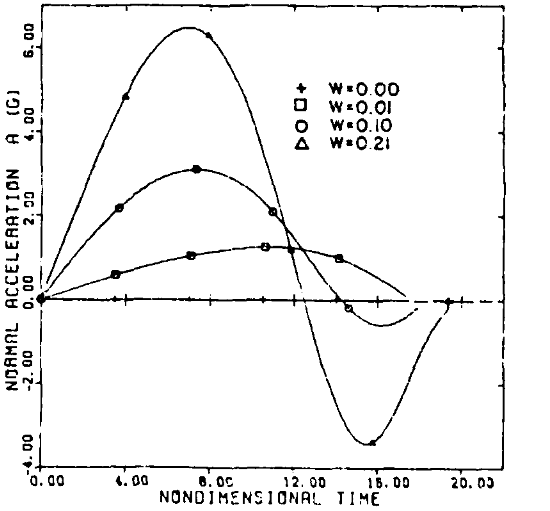


Fig. 3 Normal acceleration histories.

Table 1 Summary of numerical results[a]

W	θ_0, deg	λ_{1_0}	λ_{2_0}	Information content	Control effort	t_f, s
0.00	22.6199	0.00	0.00	5.2632	0.0000	9.2679
0.01	17.5956	−.00259	−0.00330	5.6373	.0025	9.3577
0.10	9.8100	−.02309	−.01244	5.7782	.0102	9.6648
0.21	−.8789	−.04570	−.01342	5.3825	.0268	10.4456

[a] $\rho_0 = 4.5$ (3000 ft), $v_R = 1.3$, $\phi = 30$ deg, τ_f = free, θ_f = free

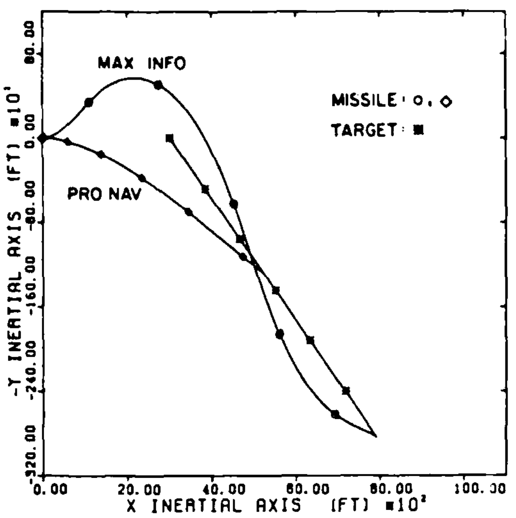


Fig. 4 Maximum information path, horizontal plane.

previous W. This procedure leads to the results presented in Table 1 and Figs. 2 and 3 for the case where the initial range is 3000 ft, the velocity ratio is $v_R = V_M/V_T = 1.3$, and the target direction is $\phi = 30$ deg.

It is noted from Table 1 that as W increases the control effort, the information content and the final time increase. The corresponding trajectories are shown in Fig. 2. For increasing W, the trajectories tend to move more toward a tail chase and oscillate back and forth behind the target. Also, the normal acceleration required to perform the maneuver, presented in Fig. 3, increases with W. For $W = 0.21$, the highest normal acceleration required is approximately 6 g.

The trajectories of Fig. 2 are similar to those obtained in Ref. 1 where the performance index is just information. For comparison purposes, the horizontal projection of the maximum information path of Ref. 1 is illustrated in Fig. 4. Note the similarity with $W = .21$ of Fig. 2.

Finally, solutions have only been obtained for values of W up to around 0.21. For $W > 0.21$, the shooting method is unable to converge to a solution. It is felt that the difficulty is caused by the minus sign in the performance index of Eq. (6). At some point, the missile can accumulate information faster than spending control to accomplish the intercept. Hence, the missile can wander around, accomplish the intercept at $t_f = \infty$, and generate $J = -\infty$.

Discussion and Conclusions

A recently-defined information index has been used to enhance the information content of minimum control-effort trajectories for the homing missile intercept problem. Optimal information-weighted trajectories have been obtained and display the desired characteristic, that is, maneuvering for the sake of increasing information content. Because of the simplicity of the model assumed here, it should be possible to compute these optimal trajectories on line and, hence, have a mechanizable guidance law for information enhancement.

Acknowledgment

This research has been sponsored by the Air Force Armament Laboratory, Eglin AFB, under contract F08635-81-R-0211, managed by Captain R.K. Liefer.

References

[1] Speyer, J.L., Hull, D.G., Tseng, C.Y. and Larson, S.W., "Estimation Enhancement by Trajectory Modulation for Homing Missiles," *Journal of Guidance, Control, and Dynamics*, Vol. 7, March-April, 1984, pp. 167-174.

[2] Bryson, A.E. and Ho., Y.C., *Applied Optimal Control*, Hemisphere Publishing Corporation, New York, 1975.

[3] Tseng, C.Y., "A Study of Maximum Information Trajectories for Homing Missile Guidance," Ph.D. Dissertation, University of Texas at Austin, Dec. 1983.

[4] Roberts, S.M. and Shipman, J.S., *Two-Point Boundary-Value Problems: Shooting Methods*, American Elsevier Publishing Company, New York, 1972.

A Stochastic Analysis of a Modified Gain Extended Kalman Filter with Applications to Estimation with Bearings Only Measurements

TAEK L. SONG, MEMBER, IEEE, AND JASON L. SPEYER, FELLOW, IEEE

Abstract—**A new globally convergent nonlinear observer, called the modified gain extended Kalman observer (MGEKO), is developed for a special class of systems. This observer structure forms the basis of a new stochastic filter mechanization called the modified gain extended Kalman filter (MGEKF). A sufficient condition for the estimation errors of the MGEKF to be exponentially bounded in the mean square is obtained. Finally, the MGEKO and the MGEKF are applied to the three-dimensional bearings-only measurement problem where the extended Kalman filter often shows erratic behavior.**

I. Introduction

The construction of implementable observers for nonlinear deterministic systems and nonlinear filters for nonlinear stochastic systems remains a challenge. With a few notable exceptions (e.g., [15]), implementation is based upon ad hoc expansions and linearization techniques. For example, in [27] an extended linear observer is developed. In [26] a nonlinear observer is implemented by augmenting the original state space by new states composed of the quadratic terms resulting from the second-order Taylor expansion of the system nonlinearities. In [24] the terms of a truncated expansion of linearly independent functions, which approximate the system nonlinearity is used to augment the state space and, thereby, construct an observer. Similarly, in stochastic nonlinear estimation problems, numerous filtering algorithms, based upon series expansions to realize approximately the conditional mean, have been suggested (see the bibliography of [8]). However, such techniques are prohibitive, in general, even for low-order dynamical systems because of the computational burden. Moreover, stability analyses of such schemes are quite rare.

As a computationally realizable and practical filter, the EKF is often used. Fortunately, there are many examples, especially in high SNR problems, where the EKF is successful in producing useful estimates. Except for a few particular cases, little is known about the properties of the estimates (i.e., stability, unbiasedness, and convergence) that it or its variants produce [5], [13]. To begin to understand some of the properties of the structure of the EKF, a nonlinear filter with a constant gain is proposed in [25] and also forms the basis for the stochastic stability analysis of [20]. The filter designed in [25] is based solely on stability considerations in a probabilistic Hilbert space. Later, in [19] and [20] the stochastic stability properties of the constant gain EKF (CGEKF) is determined in the extended inner product space M_{2_e}. In this way a certain margin of robustness is guaranteed by calculating the gain of the CGEKF from the algebraic Riccati equation (ARE) associated with a steady-state Kalman filter based on a linearized model of the actual system. However, the convergence rate of the CGEKF is found to be too slow for use in many real time estimation problems. To enhance convergence a gain-scheduling scheme is suggested [20], but the stability analysis no longer applies.

The effort described here is restricted to a special class of nonlinear functions which allows the stability analysis of [25] to be applied to an estimator where the gain changes according to an update formula that is similar to that of the EKF. This special class of nonlinear functions was motivated by the class of functions which can be manipulated into a pseudomeasurement form [1], [3], [14], [18], [29]. For deterministic systems these pseudomeasurements are linear functions of the states of the system, although the coefficient matrix is a nonlinear function of the original measurements. By using the pseudomeasurements in a linear observer structure [21], global stability can be shown. However, if the pseudomeasurement observer (PMO) is used in a noisy environment as a pseudomeasurement filter (PMF) [1], [3], [14], [18], biased estimates result. This property of the PMF is also shown in the results of Section IV. In [21] it is shown for a particular example how the EKO (the extended Kalman observer) can be manipulated into the form of the PMO. The essential difference lies only in the calculation of the observer gains. A modification of the gains of the PMO is suggested in [21] which enables the EKO to achieve performance similar to that of the PMO. This is called the modified gain EKO (MGEKO). The essential idea behind the MGEKO is that the nonlinearities be "modifiable." This notion is defined in Section II and is the central idea used in developing the structure of the estimators. This idea has some similarities with the development of the pseudomeasurement function but it is not the same. For example, the concept of modifiability also applies to nonlinear dynamic systems [22], [23].

The objective of this paper is threefold. First, the class of nonlinear modifiable functions is defined in Section II, and then a general form of the MGEKO is shown to be globally convergent under certain conditions. Secondly, the stochastic stability of the MGEKO used in the noisy environment as a modified gain EKF (MGEKF) is analyzed in Section III. Since the gains are not constant, this analysis is related but distinctly different from that of [25], [19], [20]. Note also that since the MGEKF is based on the algorithm for the EKF, the gain of the MGEKF is a function of only past measurements. This differs from the gain of the PMF which is a function of present and past measurements. Therefore, by eliminating the direct correlation of the gain and measurement noise process in the estimates of the MGEKF, the estimation bias, so prevalent in the PMF, is seen by the simulation of Section IV to be effectively eliminated. Therefore, the third objective is to apply the MGEKF to the bearings only measurement problem (BOMP) (Section IV) which has important applications in naval engagements [1], [3], [14], [18] or for homing missile engagements where passive seekers are used to track the target. Comparisons of the estimation performance of the MGEKF with the EKF and

Manuscript received March 21, 1983; revised September 19, 1983, April 30, 1984, and September 20, 1984. Paper recommended by Past Associate Editor, A. S. Willsky. This work was supported in part by Eglin AFB under Contract F08635 82-C-0090 and by the Joint Services Electronics Program under Contract F49620-77-C-0101.

The authors are with the Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin, Austin, TX 78712

0018-9286/85/1000-0940$01.00

PMF show that the MGEKF is very rapidly convergent and seemingly unbiased.

II. The Modified Gain Extended Kalman Observer (MGEKO)

In this section, a class of nonlinear functions is defined which forms the basis of the globally convergent modified gain extended Kalman observer (MGEKO). The gain structure of the MGEKO for this particular class of nonlinear system is essentially the same as that of the Kalman filter, since the error dynamics of the MGEKO are in the same form as those of a linear system. Even though this paper deals with discrete, linear system dynamics, the ideas extend to the continuous-time case and to discrete nonlinear system dynamics [22], [23].

Consider the deterministic case where the system dynamics are linear, and the measurement z_i^* is a nonlinear function of the states x_i, i.e.,

$$x_{i+1} = A_i x_i \tag{2.1}$$

$$z_i^* = h_i(x_i) \tag{2.2}$$

where $i \in Z_+$ (the nonnegative integers), $x_i \in \mathfrak{R}^n$, $z_i^* \in \mathfrak{R}^q$.

Consider the following definition of "modifiability."

Definition 1: A time-varying function $a_i: \mathfrak{R}^n \to \mathfrak{R}^q$ is *modifiable* if there exists a $q \times n$ time-varying matrix of functions $g_i: \mathfrak{R}^q \times \mathfrak{R}^n \to \mathfrak{R}^{q \times n}$ so that for any $x, \bar{x} \in \mathfrak{R}^n$ and $i \in Z_+$,

$$a_i(x) - a_i(\bar{x}) = g_i(z_i^*, \bar{x})(x - \bar{x}) \tag{2.3}$$

where $z_i^* = h_i(x)$.

Note that the difference $a_i(x) - a_i(\bar{x})$ is equal to $g_i(z_i^*, \bar{x})(x - \bar{x})$ without any approximations. Notice also that $g_i(z_i^*, \bar{x}) = g_i(h_i(x), \bar{x}) \neq g_i(h_i(\bar{x}), \bar{x})$, where this latter quantity is (if a_i is differentiable) the differential of a_i evaluated at $\bar{x}$, as used in linearization.

The MGEKO has the following structure:

$$\bar{x}_{i+1} = A_i \hat{x}_i, \tag{2.4}$$

$$\hat{x}_i = \bar{x}_i + \underline{k}_i(z_i^* - h_i(\bar{x}_i)), \tag{2.5}$$

where $\bar{x}_i$ can be interpreted as a propagated estimate at time i, $\hat{x}_i$ can be interpreted as an updated estimate, and $\underline{k}_i$ is some gain sequence that may depend on past and present data. If $h_i(x)$ is assumed modifiable, $\hat{x}_i$ in (2.5) can be rewritten as

$$\hat{x}_i = \bar{x}_i + \underline{k}_i g_i(z_i^*, \bar{x}_i)(x_i - \bar{x}_i) \tag{2.6}$$

where $g_i(z_i^*, \bar{x}_i)(x_i - \bar{x}_i) = h_i(x_i) - h_i(\bar{x}_i)$, and $g_i \in \mathfrak{R}^{q \times n}$. The error e_i in the estimates is defined as

$$\begin{aligned} e_i &\triangleq x_i - \hat{x}_i \\ &= (I - \underline{k}_i g_i(z_i^*, \bar{x}_i))\bar{e}_i \triangleq D_i \bar{e}_i \end{aligned} \tag{2.7}$$

where from (2.1) and (2.4), $\bar{e}_i$ satisfies

$$\bar{e}_i \triangleq x_i - \bar{x}_i = A_{i-1} e_{i-1}. \tag{2.8}$$

Since z_i^* is a deterministic quantity, (2.7) and (2.8) are exact and they are in the same form as that of a linear estimation system. This provides us with motivation for choosing a particular gain sequence based upon Kalman filter-type updates. Specifically,

$$m_i = A_{i-1} p_{i-1} A_{i-1}^T + Q_{i-1} \tag{2.9}$$

$$\underline{k}_i = m_i g_i(z_i^*, \bar{x}_i)^T (g_i(z_i^*, \bar{x}_i) m_i g_i(z_i^*, \bar{x}_i)^T + \gamma_i)^{-1} \tag{2.10}$$

$$p_i = (I - \underline{k}_i g_i(z_i^*, \bar{x}_i)) m_i (I - \underline{k}_i g_i(z_i^*, \bar{x}_i))^T + \underline{k}_i \gamma_i \underline{k}_i^T. \tag{2.11}$$

Note that if we had the linear estimation problem with dynamics

$$\zeta_{i+1} = A_i \zeta_i + \omega_i \tag{2.12}$$

and measurements

$$y_i = g_i(z_i^*, \bar{x}_i)\zeta_i + \upsilon_i \tag{2.13}$$

where the white noise sequences ω_i and υ_i have covariances Q_i and γ_i, respectively, then (2.9)–(2.11) would be the precise equations for the covariances m_i and p_i of the one-step predicted and filtered estimates of ζ_i, respectively. Here we are assuming that z_i^* and $\bar{x}_i$ are *known*. We will view Q_i and γ_i as design parameters and will call m_i and p_i the "pseudocovariances" of $\bar{e}_i$ and e_i, respectively.

If (A_i, g_i) is uniformly observable and $(A_i, Q_i^{1/2})$ is uniformly controllable, it can be shown that the error dynamics of (2.7) and (2.8) are globally convergent to zero by using the Lyapunov function $V_i(e_i) = e_i^T p_i^{-1} e_i$ in a way that is similar to the approach used by [17]. Note that since g_i depends upon the specific state-space trajectory, the observability condition is also trajectory dependent in general and this may be difficult to check.

Remark 1: In the next section white measurement and process noise sequences are added to the measurement (2.2) and dynamics (2.1). If the gain algorithm of (2.10) is used in this stochastic environment, biased estimates are expected, since the gain and the residual of (2.5) are directly correlated in a manner similar to the pseudomeasurement filter (PMF) as shown in the Appendix. Therefore, a gain algorithm similar to that of the EKF which ensures that the gain is a function of the *past* measurements only is recommended. However, if the measurement equation is a nonlinear function of the states, the useful relationship between the observability Gramian and the p_i matrix in (2.11) is no longer available. This discussion motivates the assumptions required in the development of the modified gain extended Kalman filter (MGEKF) presented in Section III.

III. The Modified Gain Extended Kalman Filter (MGEKF)

In this section we develop the MGEKF and study its stochastic stability. As stated in Remark 1, the gain algorithm of the MGEKF is altered from that of the MGEKO in order to reduce the biases due to direct correlations between the gain and the residual.

Consider the stochastic case where the system dynamics are linear, and the measurement z_i is a nonlinear function of the states x_i, i.e.,

$$x_i = A_{i-1} x_{i-1} + \omega_{i-1} \tag{3.1}$$

$$z_i = h_i(x_i) + \upsilon_i \triangleq z_i^* + \upsilon_i \tag{3.2}$$

where $\{\omega_i\}$ is a zero-mean independent process noise vector sequence with finite second moment

$$E\{\omega_i \omega_j^T\} Q_i \delta_{ij} \tag{3.3}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$, and where $\{\upsilon_i\}$ is a zero-mean independent measurement noise vector sequence with finite second moment

$$E\{\upsilon_i \upsilon_j^T\} = \gamma_i \delta_{ij}. \tag{3.4}$$

Furthermore, the ω_j's and υ_j's are assumed independent.

Based on Remark 1, the structure of the MGEKF is altered from that of the MGEKO. Furthermore, since z_i^* in (3.2) is not available because of the measurement noise υ_i, z_i^* is replaced by z_i in the gain formulation. If $h_i(\cdot)$ is modifiable and differentiable, the estimates of the MGEKF are obtained from the following algorithm:

$$\bar{x}_i = A_{i-1} \hat{x}_{i-1} \tag{3.5}$$

$$\hat{x}_i = \bar{x}_i + k_i(z_i - h_i(\bar{x}_i)) \tag{3.6}$$

$$m_i = A_{i-1} p_{i-1} A_{i-1}^T + Q_{i-1} \qquad (3.7)$$

$$k_i = m_i \bar{h}_{x_i}^T (\bar{h}_{x_i} m_i \bar{h}_{x_i}^T + \gamma_i)^{-1} \qquad (3.8)$$

$$p_i = (I - k_i g_i(z_i, \bar{x}_i)) m_i (I - k_i g_i(z_i, \bar{x}_i))^T + k_i \gamma_i k_i^T \qquad (3.9)$$

where $\bar{h}_{x_i} \triangleq \partial h_i(x_i)/\partial x_i|_{x_i = \bar{x}_i}$.

Note that if the predicted measurement $h_i(\bar{x}_i)$ is used in (2.10) rather than z_i^*, then the gain in (2.10) reduces to that of (3.8). In this way, the gain of the MGEKF is in the form of the EKF which ensures that the gain is only a function of the past measurements. Since the stochastic estimator (3.6) can be rewritten in a modifiable form without any approximation as

$$\hat{x}_i = \bar{x}_i + c_i(x_i - \bar{x}_i) + k_i v_i \qquad (3.10)$$

where $c_i \triangleq k_i g_i(z_i^*, \bar{x}_i)$, the error equations produced from (3.10), (3.1), and (3.5) are exact. Although c_i is not implementable, the form of (3.10) is important for our analysis of the behavior of the algorithm. Furthermore, it is critical to MGEKF performance that g_i in (3.9) be calculated using z_i. In contrast, the EKF algorithm [13] calculates g_i using $h_i(\bar{x}_i)$ in (3.9). As a final note, in [28] the MGEKF is applied to an estimation problem where part of the state vector is composed of discrete valued random variables. For this problem $g_i(z_i, \bar{x}_i)$ exists. The EKF cannot be applied to this problem since the partial derivative $\bar{h}_{x_i}$ does not exist.

A. Stochastic Stability of the Intermediate MGEKF

In order to facilitate the stability analysis of the MGEKF, as a first step we employ an unrealizable estimator, which used z_i^* to calculate the gain k_i. Although this scheme is not implementable for the noisy environment, it forms a nominal to which the implementable filter is compared. This estimator, called "the intermediate MGEKF" for convenience, is given by the following algorithm:

$$\bar{x}_i^* = A_{i-1} \hat{x}_{i-1}^* \qquad (3.11)$$

$$\begin{aligned} \hat{x}_i^* &= \bar{x}_i^* + k_i^*(z_i - h_i(\bar{x}_i^*)) = \bar{x}_i^* + k_i^* g_i(z_i^*, \bar{x}_i^*)(x_i - \bar{x}_i^*) + k_i^* v_i \\ &\triangleq \bar{x}_i^* + c_i^*(x_i - \bar{x}_i^*) + k_i^* v_i \end{aligned} \qquad (3.12)$$

$$m_i^* = A_{i-1} p_{i-1}^* A_{i-1}^T + Q_{i-1} \qquad (3.13)$$

$$k_i^* = m_i^* (\bar{h}_{x_i}^*)^T [(\bar{h}_{x_i}^*) m_i^* (\bar{h}_{x_i}^*)^T + \gamma_i]^{-1} \qquad (3.14)$$

$$p_i^* = (I - c_i^*) m_i^* (I - c_i^*)^T + k_i^* \gamma_i k_i^{*T} \qquad (3.15)$$

where the superscript * means that the estimates are obtained from the gain algorithm with $z_i^* = h_i(x_i)$ in (3.2) instead of z_i. The essential change in forming the intermediate MEGKF is that z_i^* rather than z_i is used to calculate p_i and, therefore, (3.9) is replaced by (3.15). The second equality of (3.12) uses the important modifiability characteristics of h_i to be exploited in the following analysis.

First, we consider the stability of the intermediate MGEKF by using Lyapunov's second method in the probabilistic Hilbert space L_2. The norm of a vector random variable x_i is defined as

$$\|x_i\|^2 = \int_{-\infty}^{\infty} x_i^T x_i \pi(x_i)\, dx_i, \qquad (3.16)$$

where $\pi(x_i)$ is the probability density function of x_i. Before proceeding further, the following definition is introduced.

Definition 2 [25]: A discrete stochastic process x_i is said to be exponentially bounded in mean square with exponent δ, if there exist constant $0 < \delta < 1$, $K_1 \geq 0$, and $K_2 > 0$ such that

$$\|x_i\|^2 \leq K_1 + K_2(1-\delta)^i \qquad \text{for all } i \in Z_+.$$

The errors in the estimates of the intermediate MGEKF can be written as

$$\bar{e}_i^* = x_i - \bar{x}_i^* = A_{i-1} e_{i-1}^* + \omega_{i-1} \qquad (3.17)$$

where

$$e_i^* = x_i - \hat{x}_i^* = (I - c_i^*)\bar{e}_i^* - k_i^* v_i. \qquad (3.18)$$

Define a Lyapunov function $\bar{V}_i(\bar{e}_i^*)$ as

$$\bar{V}_i(\bar{e}_i^*) = \bar{e}_i^{*T} m_i^{*-1} \bar{e}_i^*. \qquad (3.19)$$

Before stating Theorem 1, the following assumptions are needed.

Assumption 1: A_i in (3.1) is uniformly bounded and invertible.

Assumption 2: $L_i^* \triangleq (I - c_i^*)$ of (3.15) is invertible for all $i \in Z_+$.

Assumption 3: Q_i of (3.3) is uniformly bounded from below such that $Q_i \geq \alpha \cdot I > 0$ for all $i \in Z_+$.

Assumption 4: m_i^{*-1} is bounded from below by a constant matrix $c \cdot I$ such that

$$\|\bar{V}_i(\bar{e}_i^*)\| = \|\bar{e}_i^{*T} m_i^{*-1} \bar{e}_i^*\| \geq c\|\bar{e}_i^*\|^2. \qquad (3.20)$$

Remark 2: Assumptions 1 and 3 are not terribly restrictive. Note that for the MGEKO described in Section II, the uniform observability of (A_i, g_i) is sufficient to guarantee that D_i, defined in (2.7), is invertible and that m_i^{-1} is uniformly bounded from below where m_i is defined in (2.9). The corresponding conditions on L_i^* and m_i^{*-1} in Assumptions 2 and 4 are not unreasonable, although there is no such simple sufficient condition that can be checked.

Theorem 1: The errors in the estimates of (3.17) and (3.18) of the intermediate MGEKF are exponentially bounded in mean square with exponent δ under Assumptions 1–4.

The following proof of Theorem 1 is distinctly different from the proof of [25, Theorem 4] where the stability of the estimates is based on a constant gain nonlinear estimator.

Proof: Rewrite (3.17) by using (3.18) as

$$\bar{e}_{i+1}^* = A_i L_i^* \bar{e}_i^* - A_i k_i^* v_i + \omega_i. \qquad (3.21)$$

Note that v_i and L_i^*, and ω_i and L_i^* are independent, since z_i^* is not a function of ω_i and $\bar{x}_i^*$ is a function of the past measurements. Moreover, since m_{i+1}^* in (3.13) is a function of z_i^* and $\bar{x}_i^*$, m_{i+1}^* is independent of v_i and ω_i. Take the conditional expectation over $\bar{V}_{i+1}(\bar{e}_{i+1}^*) - \bar{V}_i(\bar{e}_i^*)$, given $Y_i^* = \{\bar{e}_0^*, \bar{e}_1^*, \cdots, \bar{e}_i^*\}$

$$\begin{aligned} &E_{Y_i}^*\{\bar{V}_{i+1}(\bar{e}_{i+1}^*) - \bar{V}_i(\bar{e}_i^*)\} \\ &\quad = \bar{e}_i^{*T} E_{Y_i}^*\{L_i^{*T} A_i^T m_{i+1}^{*-1} A_i L_i^* - m_i^{*-1}\}\bar{e}_i^* \\ &\qquad + E_{Y_i}^*\{\text{tr}\ [k_i^{*T} A_i^T m_{i+1}^{*-1} A_i k_i^* \gamma_i] + \text{tr}\ [m_{i+1}^{*-1} Q_i]\} \end{aligned} \qquad (3.22)$$

where $E_{Y_i}^*\{\cdot\}$ is a conditional expectation operator given Y_i^*. Note that the terms inside the tr operator become strictly positive for $Q_i \geq \alpha \cdot I > 0$ and if m_i^* obeys Assumption 4. Define

$$K_{1_i} = E_{Y_i}^*\{\text{tr}\ [\text{k}_i^{*T} A_i^T m_{i+1}^{*-1} A_i k_i^* \gamma_i] + \text{tr}\ [m_{i+1}^{*-1} Q_i]\}, \qquad (3.23)$$

then (3.22) becomes

$$\begin{aligned} &E_{Y_i}^*\{\bar{V}_{i+1}(\bar{e}_{i+1}^*) - \bar{V}_i(\bar{e}_i^*)\} \\ &\quad = K_{1_i} + \bar{e}_i^{*T} E_{Y_i}^*[\text{L}_i^{*T} A_i^T m_{i+1}^{*-1} A_i L_i^* - m_i^{*-1}\}\bar{e}_i^*. \end{aligned} \qquad (3.24)$$

The term inside the $E_{Y_i}^*$ operator on the right-hand side of (3.24) can be written after some manipulations as

$$\begin{aligned} &L_i^{*T} A_i^T m_{i+1}^{*-1} A_i L_i^* - m_i^{*-1} \\ &\quad = L_i^{*T} A_i^T (A_i p_i^* A_i^T + Q_i)^{-1} A_i L_i^* - m_i^{*-1} \\ &\quad = -s_i^{*T}(s_i^* m_i^* s_i^{*T} + \gamma_i)^{-1} s_i^* \\ &\qquad - L_i^{*T} p_i^{*-1} A_i^{-1}(Q_i^{-1} + A_i^{-T} p_i^{*-1} A_i^{-1})^{-1} A_i^{-T} p_i^{*-1} L_i^* \end{aligned} \qquad (3.25)$$

where the matrix inversion lemma [13] is used. A_i is assumed invertible, and s_i^* of the intermediate MGEKF satisfies

$$s_i^* = \gamma_i k_i^{*T} L_i^{*-T} m_i^{*-1}. \tag{3.26}$$

Therefore, from (3.25)

$$\bar{e}_i^{*T}(L_i^{*T} A_i^T m_{i+1}^{*-1} A_i L_i^* - m_i^{*-1})\bar{e}_i^* < 0. \tag{3.27}$$

Hence, there exists $0 < \rho_i \leq \beta_1 < 1$ such that

$$\bar{e}_i^{*T} L_i^{*T} A_i^T m_{i+1}^{*-1} A_i L_i^* \bar{e}_i^* = \rho_i \bar{e}_i^{*T} m_i^{*-1} \bar{e}_i^*. \tag{3.28}$$

The existence of β_1 is assured by Assumptions 2, 3, and (3.25). Now, (3.24) becomes

$$E_{Y_i}^*\{\bar{V}_{i+1}(\bar{e}_{i+1}^*) - \bar{V}_i(\bar{e}_i^*)\} \leq \underline{K}_1 - \delta_i E_{Y_i}^*\{\bar{V}_i(\bar{e}_i^*)\} \tag{3.29}$$

where $0 < \sup_{j \in Z_+} \{\underline{K}_{1_j}\} \leq \underline{K}_1 < \infty$, and $\delta_i = 1 - \rho_i$ such that $0 < \beta_2 \leq \delta_i < 1$. The boundedness of $\underline{K}_1$ is obtained from Assumptions 3, 4, and (3.14). Note that Assumption 3 implies that m_i^{*-1} is uniformly bounded from above.

By applying the nesting property of the conditional expectation to (3.29), one can obtain

$$\begin{aligned} E_{Y_0}^*\{\bar{V}_{i+1}(\bar{e}_{i+1}^*)\} &= E_{Y_0}^*\{E_{Y_i}^*\{\bar{V}_{i+1}(\bar{e}_{i+1}^*)\}\} \\ &\leq \underline{K}_1 + (1-\delta_i) E_{Y_0}^*\{E_{Y_i}^*\{\bar{V}_i(\bar{e}_i^*)\}\} \\ &= \underline{K}_1 + (1-\delta_i) E_{Y_0}^*\{\bar{V}_i(\bar{e}_i^*)\}. \end{aligned} \tag{3.30}$$

Define δ as $\delta \doteq \inf_{j \in Z_+} \{\delta_j\}$ and note that since $0 < \beta_2 \leq \delta_i < 1$, then $\delta > 0$. Applying (3.30) recursively results in

$$E_{Y_0}^*\{\bar{V}_{i+1}(\bar{e}_{i+1}^*)\} \leq \underline{K}_1 \sum_{j=0}^{i} (1-\delta)^j + (1-\delta)^{i+1} E_{Y_0}^*\{\bar{V}_0(\bar{e}_0^*)\}. \tag{3.31}$$

Use (3.19) and take an unconditional expectation over Y_0^* in (3.31). Then

$$\|\bar{e}_{i+1}^*\|^2 \leq K_1 + K_2(1-\delta)^{i+1} \tag{3.32}$$

where $K_1 < \underline{K}_1 \sum_{j=0}^{\infty} (1-\delta)^j / c = \underline{K}_1 / c\delta$ and $K_2 = E\{\bar{V}_0(\bar{e}_0^*)\}/c$.

Therefore, the exponential boundedness of the intermediate MGEKF is proved.

B. Stochastic Stability of the MGEKF

Thus far, the exponential boundedness only of the intermediate MGEKF has been proved. Now, our objective is to obtain sufficient conditions for the MGEKF to be exponentially bounded in L_2 by comparing the estimates of the MGEKF to those of the exponentially bounded intermediate MGEKF. Such conditions are found again by using Lyapunov's second method and are similar in concept to [6]. In this way the conditions for the deviation from the nominal to belong to the set of nondestabilizing deviations (terminology excerpted from [19]) are obtained.

The errors in the estimates of the MGEKF can be written from (3.1), (3.5), and (3.10) as

$$\bar{e}_{i+1} = x_{i+1} - \bar{x}_{i+1} = A_i e_i + \omega_i \tag{3.33}$$

where

$$e_i = x_i - \hat{x}_i = (I - c_i)\bar{e}_i - k_i v_i. \tag{3.34}$$

Note that the only difference between (3.33) and (3.34) for the MGEKF, and (3.17) and (3.18) of the intermediate MGEKF, results from the algorithm for the calculation of the gain of each filter. That is, the MGEKF uses z_i instead of z_i^* in the gain algorithm. Therefore, $N_{i-1} = \{v_1, v_2, \cdots, v_{i-1}\}$ contributes to the difference in the gains k_i and k_i^* since N_{i-1} affects the calculation of (3.9). Denote the gain of the MGEKF as the sum of the gain of the intermediate MGEKF and the perturbed gain due to N_{i-1}

$$k_i = k_i^* + \Delta k_i. \tag{3.35}$$

Similarly, c_i of the MGEKF is denoted as

$$c_i = c_i^* + \Delta c_i. \tag{3.36}$$

Therefore, $\bar{e}_{i+1}$ in (3.33) can be written as

$$\bar{e}_{i+1} = A_i(L_i^* - \Delta c_i)\bar{e}_i - A_i(k_i^* + \Delta k_i)v_i + \omega_i. \tag{3.37}$$

Consider the following sufficiency theorem.

Theorem 2: If Δk_i in (3.35) is bounded and Δc_i in (3.36) belongs to the set of nondestabilizing deviations such that

$$(L_i^* - \Delta c_i)^T A_i^T m_{i+1}^{*-1} A_i (L_i^* - \Delta c_i) - m_i^{*-1} < 0 \tag{3.38}$$

for all i, then under Assumptions 1–4, the error in the estimates of the MGEKF is exponentially bounded in mean square with exponent δ.

Proof: Introduce the Lyapunov function for the MGEKF as

$$\bar{V}_{i+1}(\bar{e}_{i+1}) = \bar{e}_{i+1}^T m_{i+1}^{*-1} \bar{e}_{i+1} \tag{3.39}$$

where m_{i+1}^{*-1} is bounded from below by Assumption 4. Since $L_i^* - \Delta c_i$ in (3.37) is a function of z_i^*, $\bar{x}_i$, and N_{i-1}, $L_i^* - \Delta c_i$ is independent of v_i and ω_i. Therefore, the conditional expectation of $\bar{V}_{i+1}(\bar{e}_{i+1}) - \bar{V}_i(\bar{e}_i)$ for given $Y_i = \{\bar{e}_0, \bar{e}_1, \cdots, \bar{e}_i\}$ becomes [similar to (3.22)]

$$\begin{aligned} &E_{Y_i}\{\bar{V}_{i+1}(\bar{e}_{i+1}) - \bar{V}_i(\bar{e}_i)\} \\ &\quad = \bar{e}_i^T E_{Y_i}\{(L_i^* - \Delta c_i)^T A_i^T m_{i+1}^{*-1} A_i (L_i^* - \Delta c_i) - m_i^{*-1}\}\bar{e}_i \\ &\quad\quad + E_{Y_i}\{\mathrm{tr}\ [(k_i^* + \Delta k_i)^T A_i^T m_{i+1}^{*-1} A_i (k_i^* + \Delta k_i)\gamma_i] \\ &\quad\quad + \mathrm{tr}\ [m_{i+1}^{*-1} Q_i]\}. \end{aligned} \tag{3.40}$$

If Δc_i satisfies (3.38) which is a modification of the left-hand side (3.25) in the presence of Δc_i, there exists $0 < \beta \leq \bar{\rho}_i < 1$ such that

$$\bar{e}_i^T (L_i^* - \Delta c_i)^T A_i^T m_{i+1}^{*-1} A_i (L_i^* - \Delta c_i)\bar{e}_i = \bar{\rho}_i \bar{e}_i^T m_i^{*-1} \bar{e}_i. \tag{3.41}$$

Therefore,

$$E_{Y_i}\{\bar{V}_{i+1}(\bar{e}_{i+1}) - \bar{V}_i(\bar{e}_i)\} \leq \underline{K}_1 - \delta_i E_{Y_i}\{\bar{V}_i(\bar{e}_i)\} \tag{3.42}$$

where $\delta_i = 1 - \bar{\rho}_i$, and $0 < \underline{K}_1 = \sup_{j \in Z_+} \{K_{1_j}\}$ where

$$K_{1_j} = E_{Y_i}\{\mathrm{tr}\ [(k_i^* + \Delta k_i)^T A_i^T m_{i+1}^{*-1} A_i (k_i^* + \Delta k_i)\gamma_i] + \mathrm{tr}\ [m_{i+1}^{*-1} Q_i]\}, \tag{3.43}$$

and $\underline{K}_1$ is bounded from above, since k_i^* is bounded from above by Assumptions 3, 4, and (3.14), and Δk_i is bounded from the hypothesis of the theorem. The remainde⁻ of the proof is the same as that of Theorem 1.

Remark 3: Note that the perturbations in (3.38) are due to variations in the gain calculation and not in the system nonlinearities as found in [19], [20]. Therefore, although condition (3.38) is similar in form to a condition given [19, Sect. 4.5.2], the derivation of this condition based on a time-varying algorithm rather than a constant gain algorithm is different. Although this global sufficiency condition is uncheckable analytically, it can be used as a guide for engineering evaluation. For example, a local test may be constructed for a given initial state and state estimate. The MGEKF and the intermediate MGEKF algorithms can be run for the same ensemble of measurement and process noise sequences. For each sequence the deviations of Δk_i and Δc_i can be calculated and the boundedness of Δk_i and the stability of the MGEKF via (3.38) can be assessed. This procedure was used in evaluating the performance of the bearings-only problem described in Section IV.

Remark 4: If the processes are ergodic, boundedness of $A_i(k_i^* + \Delta k_i)$ implies that the exponential stability of the MGEKF in L_2 and the finite gain stability of the MGEKF in M_{2_e} are equal by [12, Theorem 7].

Remark 5: A time-varying function $h_i:\Re^n \rightarrow \Re^q$ is called *approximately modifiable* if there exist time-varying matrices of functions $g_i:\Re^q \times \Re^n \rightarrow \Re^{q\times n}$ and $\mathcal{E}_i:\Re^n \times \Re^n \rightarrow \Re^{q\times n}$ where $\bar{e} = x - \bar{x}$ so that for any $x, \bar{x} \in \mathfrak{D} \subset \Re^n$ and $i \in Z_+$, $h_i(x) - h_i(\bar{x}) = [g_i(z_i^*, \bar{x}) + \mathcal{E}_i(x, \bar{e})]\bar{e}$ where $z_i^* = h_i(x)$ and where $\lim_{|\bar{e}|\rightarrow 0} \|\mathcal{E}_i(x, \bar{e})\| / \|g_i(z_i^*, \bar{x})\| \rightarrow 0$. The effect of the error $\mathcal{E}_i(x, \bar{e})\bar{e}$ on stability is to contribute to the deviations Δc_i and Δk_i in Theorem 2. For example, Δc_i is replaced by the deviation $\Delta \bar{c}_i \triangleq \Delta c_i + (k_i^* + \Delta k_i)\mathcal{E}_i(x, \bar{e})$ in (3.37) and also (3.38). The gain algorithm (3.7) and (3.9) uses the modifiable part g_i. The bearings-only measurements are shown to be approximately modifiable for the three-dimensional problem in Section IV and exactly modifiable for the two-dimensional case [21].

IV. Application to Bearings-Only Measurement Problem (BOMP)

In [29] a new measurement model based on a transformation of the original measurements, called pseudomeasurements, is proposed which is linear in the states of the system with a coefficient matrix composed of nonlinear functions of the original measurements. By using the pseudomeasurements, observability criteria are rigorously established for the BOMP [18]. However, as shown in [3], the resulting estimates of a linear filter structure are biased. The simulation results here demonstrate this as well. Moreover, in the Appendix, the bias in the estimate of the PMF is analyzed. By a different approach this analysis generalizes [3]. Since the PMF produces biased estimates and because of the nonlinearities of the problem, most of the studies have been conducted using various forms of the EKF for the two-dimensional BOMP (see bibliography in [10] and [16]). Recently, [2] reported successful results using the EKF for the two-dimensional BOMP formulated in a modified polar coordinate when *no* process noise is present. However, as shown by [9] through statistical consideration, and by several others through simulation studies, EKF for the BOMP formulated in a rectangular coordinate still remains a problem. Fortunately, it can be shown that the measurement equation from the two-dimensional BOMP formulated in a rectangular coordinate is both transformable to a pseudomeasurement form and a member of the class of modifiable functions (see [21, Sect. 4.2]). For the three-dimensional BOMP which is more realistic for homing missile engagement problems, the measurement equations are approximately modifiable (see Remark 5 of Section III-B for the relationship with the previous analysis).

A. System Dynamics and Pseudomeasurements for the Homing Missile Problem

The deterministic system dynamics of the missile intercept problem written in rectangular coordinates are linear

$$x_{i+1} = A_i x_i + B_i u_i \tag{4.1}$$

where the state vector x is a nine-state vector composed of three relative positions $x_R^T \triangleq [X, Y, Z]$, three relative velocities $V_R^T \triangleq [v_X, v_Y, v_Z]$, and three target accelerations $a_T^T \triangleq [a_X, a_Y, a_Z]$, and where $u^T \triangleq [a_{M_X}, a_{M_Y}, a_{M_Z}]$ is the three-dimensional missile acceleration used here as the control vector since it is assumed that the autopilot of the missile has zero-lag. Note that in implementing the estimator, the missile acceleration is assumed to be measured perfectly from the accelerometers. The control u_i is generated from the homing guidance law which is derived by using linear-quadratic theory [7] and is the basis for modern homing missile guidance. The dynamic coefficients for (4.1) are

$$A_i = \begin{bmatrix} I_3, & \Delta t I_3, & \frac{1}{\lambda^2}(e^{-\lambda\Delta t} + \lambda\Delta t - 1)I_3 \\ 0, & I_3, & \frac{1}{\lambda}(1 - e^{-\lambda\Delta t})I_3 \\ 0, & 0, & e^{-\lambda\Delta t}I_3 \end{bmatrix}, \quad B_i = \begin{bmatrix} -(\Delta t^2/2)I_3 \\ -\Delta t I_3 \\ 0 \end{bmatrix} \tag{4.2}$$

where I_n is an $n \times n$ identity matrix, Δt is the time interval between measurements, and λ is determined from the bandwidth of the target acceleration assumed as a band-limited colored noise process. Note that A_i in (4.2) satisfies Assumption 1 in Section III-A.

The intercept geometry, measurement angles, and relative range are given in Fig. 1. The azimuth and elevation angle measurements for the three-dimensional BOMP can be written as (2.2) where z_i^* satisfies

$$\begin{bmatrix} az \\ el \end{bmatrix} = h_i(x_i) = \begin{bmatrix} \tan^{-1}\frac{Y_i}{X_i} \\ \tan^{-1} -Z_i/(X_i^2 + Y_i^2)^{1/2} \end{bmatrix}. \tag{4.3}$$

By using simple trigonometric identities, the two measurements of (4.3) are manipulated into the following pseudomeasurements $(y_i^T(z_i^*) \triangleq [y_{1_i}, y_{2_i}])$:

$$\begin{bmatrix} y_{1_i} \\ y_{2_i} \end{bmatrix} = \begin{bmatrix} \sin az, & -\cos az, & 0, & \cdots 0 \\ \sin el \cos az, & \sin el \sin az, & \cos el, & 0 \cdots 0 \end{bmatrix} x_i$$

$$\triangleq H_i(z_i^*)x_i = [0, 0]^T. \tag{4.4}$$

For the stochastic version of the system, the dynamics (4.1) of the BOMP formulated in rectangular coordinates are corrupted by additive process noise ω_i where ω_i is zero mean white noise with second moment $Q_i = E\{\omega_i \omega_i^T\}$. Since the target model is assumed as a continuous first-order Gauss–Markov process along each axis, Q_i is obtained as the resulting discrete process noise variance where Q^* is the assumed power spectral density for the continuous input process. Q_i is written as

$$Q_i = \int_0^{\Delta t} \phi(\Delta t - \tau)CQ^*C^T\phi^T(\Delta t - \tau)\, d\tau \tag{4.5}$$

where $\phi(\Delta t - \tau)$ is A_i with $\Delta t - \tau$ replacing Δt, and $C = [0:0:I_3]^T$. Note that Q_i of (4.5) satisfies Assumption 3 of Section III-A. Similarly, in the noisy environment, zero mean white noise process $v_i = [v_{1_i}, v_{2_i}]^T$ is added to $h_i(x_i)$ in (4.3) as (3.2) to form the noisy measurement equations where v_{1_i} and v_{2_i} are azimuth and elevation angle measurement noise, respectively, and $E\{v_i v_i^T\} = \gamma_i \in \Re^{2\times 2}$.

The pseudomeasurement observer [21] can be extended to the noisy environment. In that case, the pseudomeasurements [i.e., (4.4)] are corrupted by the state-dependent noises. These state-dependent noises together with the gain structure of the pseudomeasurement filter (PMF) [1], [21] cause the biases in the estimates. In the Appendix, the biases in the estimate of the PMF are analyzed by using the innovation processes for the case where the process noise is included, whereas in [3], the bias analysis for the no process noise case is studied by a batch estimation scheme. Since the noise variance R_i corresponding to the pseudomeasure-

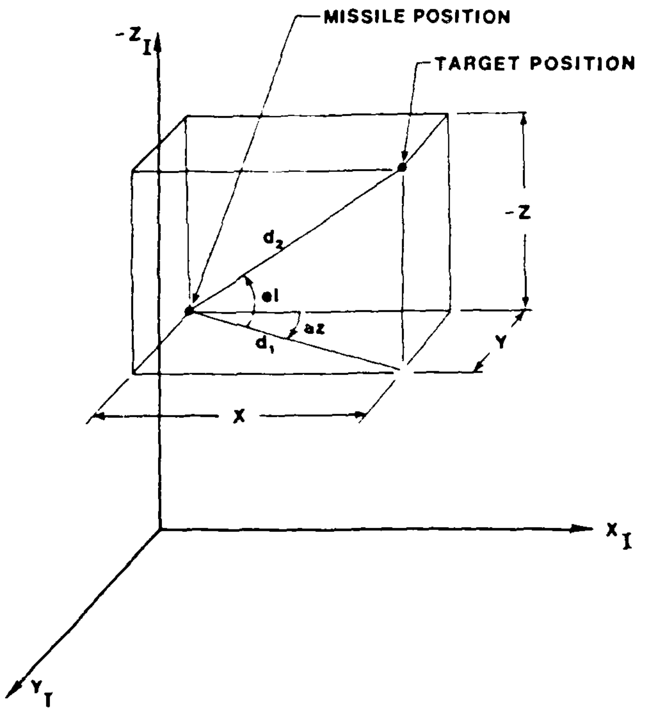


Fig. 1. Intercept geometry and measurement angles.

ment, as shown in the Appendix, is associated with the state-dependent noise (A.2) and γ_i is the noise variance of the original angle measurement. γ_i and R_i are related by

$$\gamma_i = D_i^{-1} R_i D_i^{-T} \tag{4.6}$$

where the range matrix D_i is

$$D_i = \begin{bmatrix} \sqrt{X_i^2 + Y_i^2} & 0 \\ 0 & \sqrt{X_i^2 + Y_i^2 + Z_i^2} \end{bmatrix}. \tag{4.7}$$

Since the actual state x_i is not available, γ_i is approximated as

$$\gamma_i \cong \bar{D}_i^{-1} R_i \bar{D}_i^{-T} \tag{4.8}$$

where $\bar{D}_i$ is D_i calculated using the state estimates $\bar{x}_i$ of the estimation algorithm.

B. Modifiability of $h_i(x_i)$

In this section it is shown that the measurement equation of the two-dimensional BOMP is a member of the class of modifiable functions, and $h_i(x_i)$ in (4.3) for the three-dimensional BOMP is approximately modifiable. The measurement residual in (2.5) is now manipulated for the BOMP as

$$h_i(x_i) - h_i(\bar{x}_i) = \begin{bmatrix} \tan^{-1}\left[\dfrac{Y_i}{X_i}\right] - \tan^{-1}\left[\dfrac{\bar{Y}_i}{\bar{X}_i}\right] \\ \tan^{-1}\left[\dfrac{-Z_i}{d_{1_i}}\right] - \tan^{-1}\left[\dfrac{-\bar{Z}_i}{\bar{d}_{1_i}}\right] \end{bmatrix}$$

$$\triangleq \begin{bmatrix} \tan^{-1} \alpha_i \\ \tan^{-1} \beta_i \end{bmatrix} \tag{4.9}$$

where $d_{1_i} = \sqrt{X_i^2 + Y_i^2}$. Let $\underline{D}_{1_i} \triangleq d_{1_i}/(X_i\bar{X}_i + Y_i\bar{Y}_i)$ and $\underline{D}_{2_i} \triangleq d_{2_i}/(d_{1_i}\bar{d}_{1_i} + Z_i\bar{Z}_i)$ where $d_{2_i} = \sqrt{X_i^2 + Y_i^2 + Z_i^2}$. Furthermore, note that $\sin az = Y_i/d_{1_i}$, $\cos az = X_i/d_{1_i}$, $\sin el = -Z_i/d_{2_i}$, and $\cos el = d_{1_i}/d_{2_i}$. Therefore, $\underline{D}_{1_i} = \underline{D}_{1_i}(z_i^*, \bar{x}_i)$ and $\underline{D}_{2_i} = \underline{D}_{2_i}(z_i^*, \bar{x}_i)$. Hence, α_i and β_i in (4.9) satisfy

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \underline{D}_{1_i}, & 0 \\ 0, & \underline{D}_{2_i} \end{bmatrix} \begin{bmatrix} \sin az\bar{X}_i - \cos az\bar{Y}_i \\ \sin el\bar{d}_{1_i} + \cos el\bar{Z}_i \end{bmatrix}. \tag{4.10}$$

If we add and then subtract $\underline{D}_{2_i} \sin el/\underline{D}_{1_i}$ to β_i of (4.10), then

$$\beta_i = \underline{D}_{2_i}(\sin el \cos az\bar{X}_i + \sin el \sin az\bar{Y}_i + \cos el\bar{Z}_i + \sin el\,(\bar{d}_{1_i} - 1/\underline{D}_{1_i})). \tag{4.11}$$

The last term in (4.11), which we denote by $\delta_i(x_i, \bar{e}_i)$, can be bounded as

$$\begin{aligned} &\delta_i(x_i, \bar{e}_i)/f_i(x_i, \bar{x}_i) \\ &= d_{1_i}\bar{d}_{1_i} - (X_i\bar{X}_i + Y_i\bar{Y}_i) \le (X_i - \bar{X}_i)^2 \\ &\quad + (Y_i - \bar{Y}_i)^2 \le \bar{e}_i^T \bar{e}_i \end{aligned} \tag{4.12}$$

where $f_i(x_i, \bar{x}_i) = \tan el/(d_{1_i}\bar{d}_{1_i} + Z_i\bar{Z}_i)$. This bound is used to show that the elevation angle measurement is approximately modifiable. By using $H(z_i^*)$ given in (4.4), (4.10) becomes

$$\begin{aligned} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} &= \begin{bmatrix} \alpha_i \\ \bar{\beta}_i \end{bmatrix} + \begin{bmatrix} 0 \\ \delta_i(x_i, \bar{e}_i) \end{bmatrix} \\ &= \begin{bmatrix} \underline{D}_{1_i}, & 0 \\ 0, & \underline{D}_{2_i} \end{bmatrix} H_i(z_i^*)\bar{x}_i + \begin{bmatrix} 0 \\ \delta_i(x_i, \bar{e}_i) \end{bmatrix}. \end{aligned} \tag{4.13}$$

By using $[\alpha_i, \beta_i]^T$ of (4.13) and (4.4), the measurement residual in (4.9) can be written as

$$h_i(x_i) - h_i(\bar{x}_i) = -E_i H_i(z_i^*)(x_i - \bar{x}_i) + E_i[0, \delta_i(x_i, \bar{e}_i)/\underline{D}_{2_i}]^T \tag{4.14}$$

where

$$E_i \triangleq \begin{bmatrix} (\underline{D}_{1_i} \tan^{-1} \alpha_i)/\alpha_i, & 0 \\ 0, & (\underline{D}_{2_i} \tan^{-1} \beta_i)/\beta_i \end{bmatrix}. \tag{4.15}$$

If (4.13) is used explicitly in (4.15), then (4.14) takes on the approximately modifiable form as defined in Remark 5 where the modifiable part is $g_i(z_i^*, \bar{x}_i)\bar{e}_i = -\bar{E}_i(z_i^*, \bar{x}_i)H_i(z_i^*)\bar{e}_i$ where $\bar{E}_i$ is E_i evaluated with $\bar{\beta}_i$ rather β_i, and the remainder term becomes $-\{(E_i - \bar{E}_i)H_i(z_i^*)\bar{e}_i - E_i[0, \delta_i(x_i, \bar{e}_i)/\underline{D}_{2_i}]^T\}$. Note that the nonzero element of $E_i - \bar{E}_i$ can be written as an explicit function of δ_i as $(\bar{\beta}_i \tan^{-1}(\delta_i/(1 + \beta_i\bar{\beta}_i)) - \delta_i \tan^{-1}\bar{\beta}_i)/\underline{D}_{2_i}/\beta_i\bar{\beta}_i$. This remainder has the property ascribed to the term $\mathcal{E}_i(x_i, \bar{e}_i)\bar{e}_i$ in Remark 5 in an appropriate region $\mathfrak{D}$ since as $\|\bar{e}_i\| \to 0$ this remainder is proportional to $\delta_i(x_i, \bar{e}_i)$ which has the quadratic error bound given in (4.12). By inspection of $f_i(x_i, \bar{x}_i)$, the region $\mathfrak{D}$ appears to be all of $x, \bar{x} \in \mathcal{R}^9$ except for $1/(1/d_{1_i} + 1/\bar{d}_{2_i} + 1/|\cos(el - \overline{el})|) \le \epsilon$ where ϵ is some small positive number, and $\bar{d}_{2_i}$ and $\overline{el}$ are the values of d_{2_i} and el using the state estimates.

For the noisy environment $g_i(z_i, \bar{x}_i)$ is used to calculate the gain of the MGEKF, where z_i in this case is of the form given in (3.2). Although the measurement el is approximately modifiable through β_i of (4.13), the measurement az belongs to the class of modifiable functions [21].

C. Simulation Results and Comparisons

The results given in [21] are for the two-dimensional BOMP and noiseless environment. The simulation study given here considers the three-dimensional BOMP where both the noisy and noiseless environments are included. The launch scenario is given by

$$X_0 = 3500 \text{ ft}, \ Y_0 = 1500 \text{ ft}, \ Z_0 = 1000 \text{ ft}$$

$$v_{X_0} = -1100 \text{ ft/s}, \ v_{Y_0} = -150 \text{ ft/s}, \ v_{Z_0} = -50 \text{ ft/s}$$

$$a_{T_X} = 10 \text{ ft/s}^2, \ a_{T_Y} = 10 \text{ ft/s}^2, \ a_{T_Z} = 10 \text{ ft/s}^2$$

and the initial estimates $\hat{x}_0$ of the state x_0 are assumed to be

$$\hat{X}_0 = 3000 \text{ ft}, \ \hat{Y}_0 = 1200 \text{ ft}, \ \hat{Z}_0 = 800 \text{ ft}$$

$$\hat{v}_{X_0} = -950 \text{ ft/s}, \ \hat{v}_{Y_0} = -100 \text{ ft/s}, \ \hat{v}_{Z_0} = -100 \text{ ft/s}$$

$$\hat{a}_{T_{X_0}} = 0 \text{ ft/s}^2, \ \hat{a}_{T_{Y_0}} = 0 \text{ ft/s}^2, \ \hat{a}_{T_{Z_0}} = 0 \text{ ft/s}^2.$$

The filters (or observers) are initialized with a diagonal P_0 matrix where

$$P_0 = \left[\begin{array}{c:c} 10^4 I_6 & 0 \\ \hdashline 0 & 10^2 I_3 \end{array}\right]. \tag{4.16}$$

The noise variance corresponding to the pseudomeasurement R_i (see the Appendix) is assumed constant such that $R_i = 0.1 \ I_2$ and the variance of the process noise Q_i in (4.5) is found using the power spectral density $Q^* = 0.1 \ I_3$. For the system dynamics, $\lambda = 1$ is used, and the sampling rate used in the simulation is 50 Hz.

The value of weighting between control effort and terminal miss in the quadratic cost criterion to generate the guidance command u_i [7] is 10^{-4}. Clearly, in practice the control law is mechanized by using the estimated value of the states rather than the states themselves as implemented in our simulation. However, since the emphasis here is on filter performance, the guidance law is only used to establish trajectories from which the observers and filters are tested and compared.

The performance of the observers (PMO, EKO, MGEKO) and filters (PMF, EKF, MGEKF) is measured here by using the histories of errors in the estimates. By comparing the error histories of the observers to those of the filters, the biases, so prevalent in the PMF, can be demonstrated to be effectively reduced in the MGEKF.

As shown in Figs. 2–4 for the deterministic system, the responses of the PMO and the MGEKO are quite similar, although the PMO performs a bit slower than the MGEKO. If the initial errors were reversed in sign, then the PMO performs a bit faster than the MGEKO. Note that the responses of the MGEKO in this case are obtained by using the same algorithm as the MGEKF. Therefore, the gain k_i is a function of the past measurements. This is done to compare the performance of the observer and filter with the same structure but used in different noise environments. Clearly, the EKO performs poorly for these relatively large initial errors.

For the stochastic environment, the results of 100 runs of Monte Carlo simulation are presented in Figs. 5–7. The error in the range estimate at a specific time i in Fig. 5, for example, is plotted by using the value $\sqrt{E\{e_{X_i}\}^2 + E\{e_{Y_i}\}^2 + E\{e_{X_i}\}^2}$ where $E\{e_{X_i}\}$ is the averaged value of the error in the estimate of the X_i over 100 runs of Monte Carlo simulation. Similar rms-type quantities are plotted in the other figures. Figs. 5–7 show that the estimates of the PMF are biased away from the deterministic responses of Figs. 2–4. While the MGEKF shows good tracking performance, the EKF remains poor. When the initial errors are small, the three observers for the deterministic system perform equally. However, for the noisy environment, the EKF and the MGEKF perform similarly, while the PMF still shows biases which are quite affected by P_0 as shown in (A.17). This is particularly true for the short-time engagement problem of the homing missile. Therefore, at least for this scenario, very

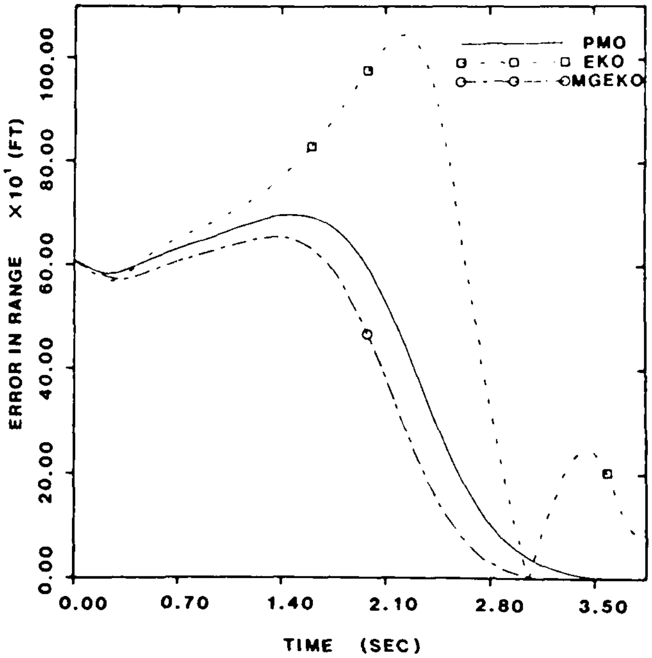


Fig. 2. Errors in range estimates of the PMO, EKO, and MGEKO for noiseless environment.

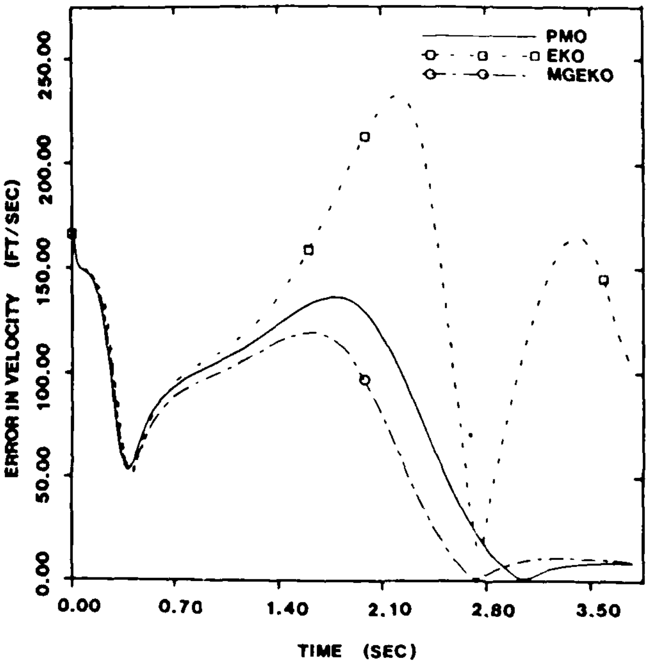


Fig. 3. Errors in velocity estimates of the PMO, EKO, and MGEKO for noiseless environment.

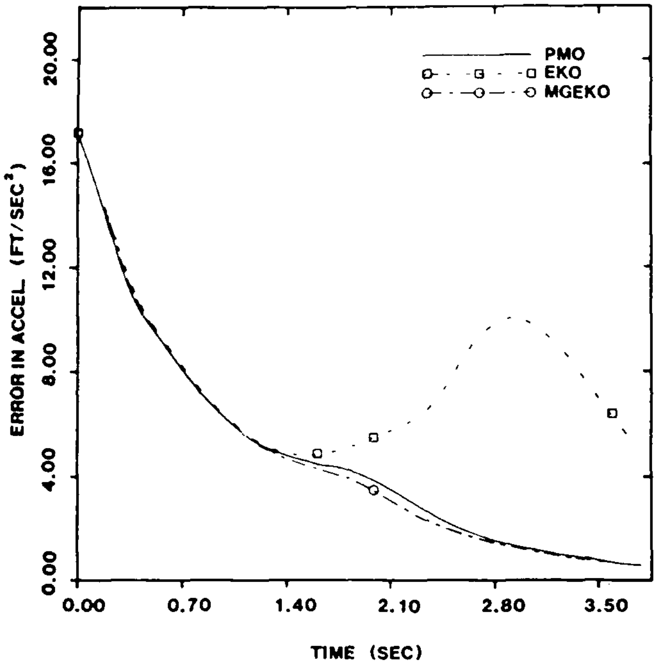


Fig. 4. Errors in acceleration estimates of the PMO, EKO, and MGEKO for noiseless environment.

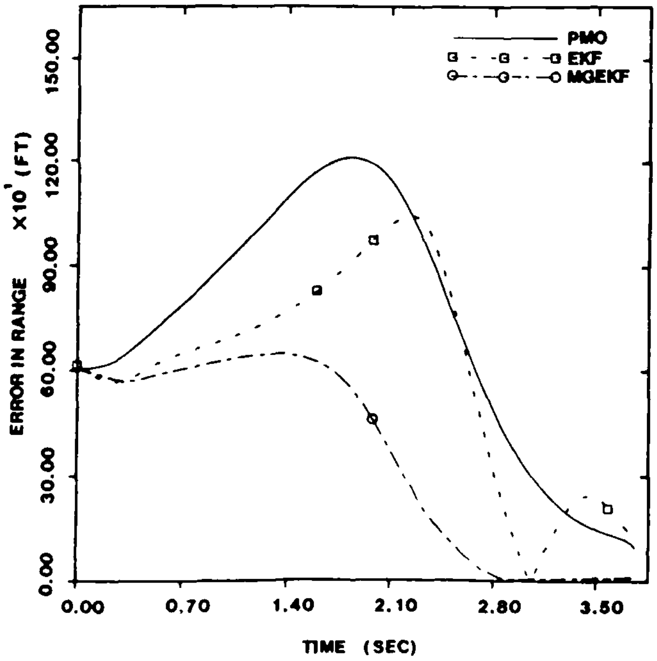


Fig 5. Errors in range estimates of the PMF, EKF, and MGEKF for noisy environment.

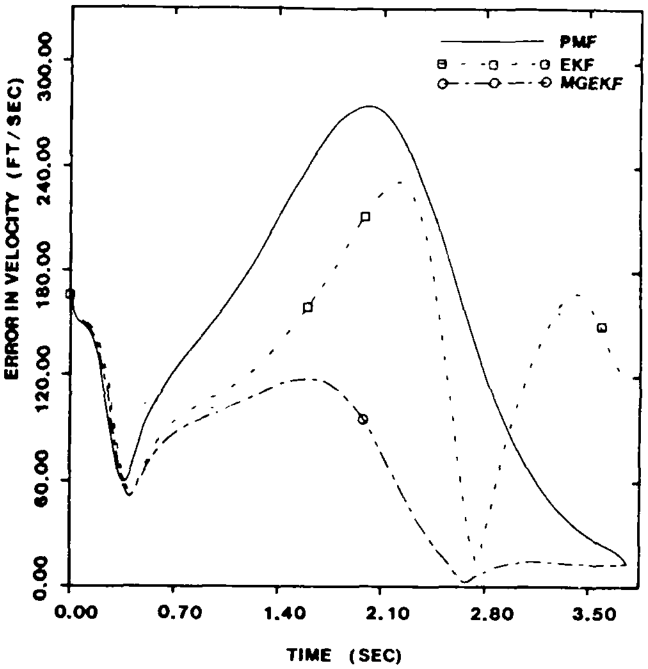


Fig. 6. Errors in velocity estimates of the PMF, EKF, and MGEKF for noisy environment.

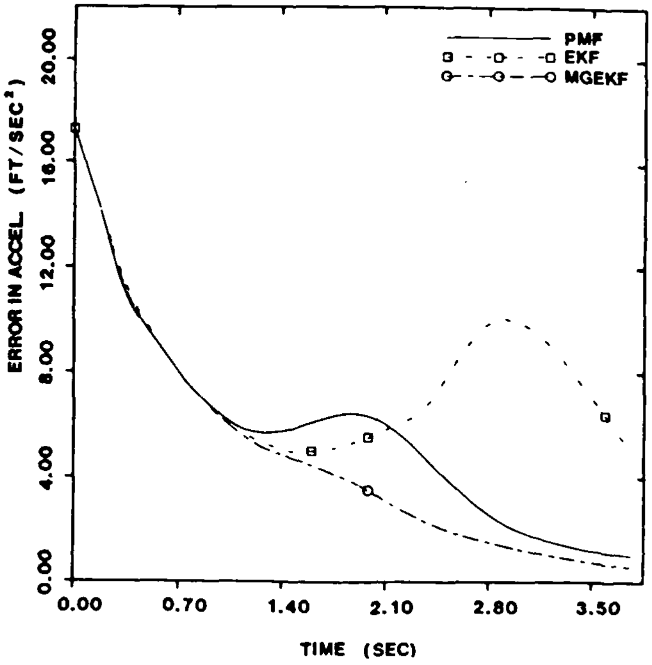


Fig. 7. Errors in acceleration estimates of the PMF, EKF, and MGEKF for noisy environment.

accurate estimates of the initial states are critical for the PMF and the EKF, while the MGEKF performs well under all conditions. To determine the deviations Δk and Δc starting from the given initial state and state estimate, the intermediate MGEKF was run using the same ensemble of measurement and process noise sequences as used in evaluating the MGEKF. Over the ensemble of realizations of these deviations, the boundedness of Δk_i required by Theorem 2 was never even close to being violated and the stochastic stability of the MGEKF by condition (3.38) was always satisfied.

V. Conclusions

A new observer, called the MGEKO for a class of nonlinear functions called modifiable functions, is designed such that it is globally stable. A stability analysis of the MGEKF is performed by introducing as an intermediate step a nominal filter called the intermediate MGEKF. In Theorem 1, the intermediate MGEKF is shown to be globally stable in the probabilistic Hilbert space L_2. Sufficient conditions for the MGEKF to be globally stable are found in a similar way to that of the intermediate MGEKF, but a condition on the allowable deviations to be nondestabilizing is required in Theorem 2.

The MGEKO and the MGEKF are applied to the three-dimensional BOMP formulated in rectangular coordinates which by Remark 5 is approximately modifiable. For the three-dimensional deterministic formulation of the BOMP, the MGEKO performs in the simulation almost identically to the PMO. However, the EKO can only estimate the states when the initial errors in the estimates are small. For the stochastic formulation, these observer structures are retained. While the estimates of the PMF are biased, the EKF and the MGEKF show seemingly unbiased characteristics in the simulation. However, the EKF appears only stable in the small. Although the simulation results indicate that the MGEKF is stochastically stable, checking the conditions of Theorem 2, with respect to an ensemble of measurement and process noise sequences starting from a given initial state and state estimate, produces additional evidence of stochastic stability.

Appendix

The objective of this Appendix is to generalize the results of [3] and to show the structure of the PMF used in the simulation and its inherent biased characteristics in a stochastic environment.

The pseudomeasurement $y_i(z_i^*)$ in (4.4) is changed to the following form in the noisy environment:

$$y_i(z_i) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = H_i(z_i)x_i + \mu_i(x_i, v_i) \tag{A.1}$$

where the noise corrupted z_i is from (3.2), and the state dependent noise μ_i is

$$\mu_i(x_i, v_i) = -[(X_i^2 + Y_i^2)^{1/2}v_{1_i},\ (X_i^2 + Y_i^2 + Z_i^2)^{1/2}v_{2_i}]^T \tag{A.2}$$

v_{1_i} and v_{2_i} are the original angle measurement noises defined in Section IV-A.

The algorithm of the pseudomeasurement filter (PMF) [1], [21], similar to the Kalman filter, satisfies

$$\bar{x}_{i+1} = A_i\hat{x}_i + B_iu_i \tag{A.3}$$

$$\hat{x}_i = \bar{x}_i + K_i(y_i(z_i) - H_i(z_i)\bar{x}_i) \tag{A.4}$$

$$P_i = (I - K_iH_i(z_i))M_i \tag{A.5}$$

$$K_i = M_iH_i(z_i)^T(H_i(z_i)M_iH_i(z_i)^T + R_i)^{-1} \tag{A.6}$$

$$M_{i+1} = A_iP_iA_i^T + Q_i. \tag{A.7}$$

Therefore, the error in the estimates e_i of the PMF can be written, by using (4.6), (A.3), and (A.4), as

$$e_i = \psi_{i,0}e_0 + \sum_{j=0}^{i-1} \psi_{i,j+1}A^{-1}\omega_j - \sum_{j=1}^{i} \psi_{i,j}k_j\mu_j \tag{A.8}$$

where

$$\psi_{i,j} = \prod_{k=j}^{i} (I - K_kH_k)A \triangleq \prod_{k=j}^{i} L_kA. \tag{A.9}$$

Since A_i in (4.2) is a constant matrix, the subscript i is suppressed. In order to understand the biased behavior of the PMF estimates, a closed-form expression for the error e_i is to be determined.

Introduce the transformation τ_i as

$$\begin{bmatrix} \mathcal{U}_i \\ \mathcal{V}_i \end{bmatrix} = \tau_i \begin{bmatrix} \mathcal{U}_{i-1} \\ \mathcal{V}_{i-1} \end{bmatrix} = \begin{bmatrix} A^{-T} + H_i^TR_i^{-1}H_iQA^{-T} & H_i^TR_i^{-1}H_iA \\ QA^{-T} & A \end{bmatrix} \begin{bmatrix} \mathcal{U}_{i-1} \\ \mathcal{V}_{i-1} \end{bmatrix} \tag{A.10}$$

where $\mathcal{U}_i \in \mathcal{R}^{n\times n}$, $\mathcal{V}_i \in \mathcal{R}^{n\times n}$. Let $\mathcal{U}_0 = P_0^{-1}$, and $\mathcal{V}_0 = I$. Then,

$$\begin{bmatrix} \mathcal{U}_i \\ \mathcal{V}_i \end{bmatrix} = \tau_i\tau_{i-1} \cdots \tau_1\tau_0 \begin{bmatrix} \mathcal{U}_0 \\ \mathcal{V}_0 \end{bmatrix} \triangleq \begin{bmatrix} \psi_{\tau_{1,1}}(i, 0) & \psi_{\tau_{1,2}}(i, 0) \\ \psi_{\tau_{2,1}}(i, 0) & \psi_{\tau_{2,2}}(i, 0) \end{bmatrix} \begin{bmatrix} P_0^{-1} \\ I \end{bmatrix}. \tag{A.11}$$

Since the solution to the discrete Riccati equation is

$$P_i = [\psi_{\tau_{2,1}}(i, 0)P_0^{-1} + \psi_{\tau_{2,2}}(i, 0)][\psi_{\tau_{1,1}}(i, 0)P_0^{-1} + \psi_{\tau_{1,2}}(i, 0)]^{-1} \tag{A.12}$$

then $P_i = \mathcal{V}_i\mathcal{U}_i^{-1}$. L_iA in (A.9) can be written in terms of $\mathcal{U}_i$, after some manipulations, as

$$L_iA = \mathcal{U}_i^{-T}\mathcal{U}_{i-1}^T. \tag{A.13}$$

Now, e_i due to μ_j's can be written, by using (A.1), (A.8), (A.9), and (A.13), as

$$\begin{aligned}(e_i)_\mu &= P_iH_i^TR_i^{-1}H_ix_i + L_iAP_{i-1}H_{i-1}^TR_{i-1}^{-1}H_{i-1}x_{i-1} + \cdots \\ &= P_i\mathcal{V}_i^{-T}\sum_{j=1}^{i} \mathcal{V}_j^TH_j^TR_j^{-1}H_jx_j.\end{aligned} \tag{A.14}$$

Similarly, the e_i due to ω_j and the e_i due to e_0 can also be written as

$$(e_i)_\omega = P_i\mathcal{V}_i^{-T}\sum_{j=0}^{i-1} \mathcal{U}_j^TA^{-1}\omega_j \tag{A.15}$$

and

$$(e_i)e_0 = P_i\mathcal{V}_i^{-T}P_0^{-1}e_0. \tag{A.16}$$

Therefore,

$$e_i = P_i\mathcal{V}_i^{-T}\left(P_0^{-1}e_0 + \sum_{j=0}^{i-1} \mathcal{U}_j^TA^{-1}\omega_j + \sum_{j=1}^{i} \mathcal{V}_j^TH_j^TR_j^{-1}x_j\right). \tag{A.17}$$

Because of the last term of (A.17) which is the error due to the state dependent noise μ, $E[e_i]$ is nonzero.

ACKNOWLEDGMENT

The authors wish to thank the associate editor whose many insightful suggestions have greatly enhanced this paper.

REFERENCES

[1] V. J. Aidala, "Behavior of the Kalman filter applied to bearing-only target motion analysis," *Advances in Passive Tracking*, Vol. 1, Naval Postgraduate School, MPS-62 TS-77071, May 1977.

[2] V. J. Aidala and S. E. Hammel, "Utilization of modified polar coordinates for bearing-only tracking," *IEEE Trans. Automat. Contr.*, vol. AC-28, Mar. 1983.

[3] V. J. Aidala and S. C. Nardone, "Biased estimation properties of the pseudolinear tracking filter," *IEEE Trans. Aerospace Electron. Syst.*, vol. AES-18, July 1982.

[4] B. D. O. Anderson and J. B. Moore, "Detectability and stabilizability of time-varying discrete-time linear systems," *SIAM J. Contr. Optimiz.*, vol. 19, Jan. 1981.

[5] M. Athans, R. P. Wishner, and A. Bertolini, "Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements," *IEEE Trans. Automat. Contr.*, vol. AC-13, Oct. 1968.

[6] S. Barnett and C. Storey, "Some results on the sensitivity and synthesis of asymptotically stable linear and nonlinear systems," *Automatica*, vol. 4, 1968.

[7] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*. Waltham, MA: Blaisdell, 1969.

[8] R. S. Bucy, C. Hecht, and K. D. Senne, "An engineer's guide to building nonlinear filters, Volume I," SRL-TR-72-0004, May 1972.

[9] S. I. Chou, "Some drawbacks of extended Kalman filters in ASW passive angle tracking," *Advances in Passive Tracking*, Vol. I, Naval Postgraduate School, MPS-62 TS-77071, May 1977.

[10] S. Fagerlund, "Target tracking based on bearing only measurements," Mass. Inst. Technol., LIDS-R-1003, June 1980.

[11] A. E. Gelb, *Applied Optimal Estimation*. Cambridge, MA: M.I.T. Press, 1974.

[12] D. J. Hill and P. J. Moylan, "Connections between finite-gain and asymptotic stability," *IEEE Trans. Automat. Contr.*, vol. AC-25, Oct. 1980.

[13] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.

[14] A. G. Lindgren and K. F. Gong, "Position and velocity estimation via bearing observations," *IEEE Trans. Aerospace Electron. Syst.*, vol. AES-14, July 1978.

[15] S. I. Marcus and A. S. Willsky, "Algebraic structure and finite dimensional nonlinear estimation," *SIAM J. Math. Anal.*, Apr. 1978.

[16] R. K. Mehra, "A comparison of several nonlinear filters for reentry vehicle tracking," *IEEE Trans. Automat. Contr.*, vol. AC-16, Aug. 1971.

[17] J. B. Moore and B. D. O. Anderson, "Coping with singular transition matrices in estimation and control stability theory," *Int. J. Contr.*, vol. 3, 1980.

[18] S. C. Nardone and V. J. Aidala, "Observability criteria for bearing-only target motion analysis," *IEEE Trans. Aerospace Electron. Syst.*, vol. AES-17, Mar. 1981.

[19] M. G. Safonov, *Stability and Robustness of Multivariable Feedback Systems*. Cambridge, MA: M.I.T. Press, 1980.

[20] M. G. Safonov and M. Athans, "Robustness and computational aspects of nonlinear stochastic estimators and regulators," *IEEE Trans. Automat. Contr.*, vol. AC-23, Aug. 1978.

[21] J. L. Speyer and T. L. Song, "A comparison between pseudomeasurement and extended Kalman observers," *Proc. 20th IEEE Conf. Decision Contr.*, San Diego, CA, Dec. 1981.

[22] T. L. Song, "A stochastic analysis of a modified gain extended Kalman filter," Ph.D. dissertation, Univ. Texas at Austin, 1983.

[23] T. L. Song and J. L. Speyer, "The modified gain extended Kalman filter and parameter identification in linear systems," *Automatica*, to be published.

[24] H. Takata, "Transformation of a nonlinear system into an augmented linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-24, Oct. 1979.

[25] T. J. Tarn and Y. Rasis, "Observers for nonlinear stochastic systems," *IEEE Trans. Automat. Contr.*, vol. AC-21, Aug. 1976.

[26] S. Tsuji, H. Takata, R. Ueda, and S. Takata, "Second-order observer for nonlinear systems from discrete noiseless measurements," *IEEE Trans. Automat. Contr.*, vol. AC-22, Feb. 1977.

[27] R. Ueda, H. Takata, S. Nakagaki, and S. Takata, "On the estimation of transient state of power system by discrete nonlinear observers," *IEEE Trans. Power Apparatus Syst.*, vol. PAS-94, Nov./Dec. 1975.

[28] E. K. Westwood, "Filtering algorithms for the linear estimation problem with switching parameters," M.S. thesis, Dep. Elec. Eng., Univ. Texas at Austin, May 1984.

[29] D. W. Whitcombe, "Pseudo state measurements applied to recursive nonlinear filtering," in *Proc. 3rd Symp. Nonlinear Estimation Theory and Its Appl.*, San Diego, CA, Sept. 1972.

Taek L. Song (S'83-M'83) was born in Andong, Korea, on August 28, 1952. He received the B.S. degree in nuclear engineering from Seoul National University, Seoul, Korea, in 1974, and both the M.S. and Ph.D. degrees in aerospace engineering from the University of Texas, Austin, in 1981 and 1983, respectively.

At present he is with the Agency for Defense Development, Daejeon, Korea, where from 1974 to 1980 he worked as a Research Engineer in the field of guidance and control of missile systems. From 1980 to 1983, he was a Research Assistant at the University of Texas, Austin, and in 1984 he became a Lecturer at the same university. His research interests are estimation theory, adaptive control theory, and multivariable system design.

Dr. Song is a member of Sigma Gamma Tau.

Jason L. Speyer (M'71-SM'82-F'85) was born in Boston, MA, in 1938. He received the B.S. degree in aeronautics and astronautics from the Massachusetts Institute of Technology, Cambridge, in 1960, and the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA, in 1968.

His industrial experience includes research at Boeing, Raytheon, Analytical Mechanics Associates, and Charles Stark Draper Laboratory. At present, he is Harry H. Power Professor in Engineering in the Department of Aerospace Engineering and Engineering Mechanics, University of Texas, Austin. He recently spent a research leave as a Lady Davis Visiting Professor at the Technion—Israel Institute of Technology, Haifa, Israel.

Dr. Speyer has been Associate Editor for Technical Notes and Correspondence and for Stochastic Control of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and is presently an elected member of the Board of Governors of the IEEE Control Systems Society and Chairman of the Technical Committee on Aerospace Controls. He was also Associate Editor for the *AIAA Journal of Spacecraft and Rockets* and the *AIAA Journal of Guidance and Control*. He is a Fellow of the American Institute of Aeronautics and Astronautics.

Automatica, Vol. 22, No. 1, pp. 59-75, 1986
Printed in Great Britain

0005-1098/86 $3.00 + 0.00
Pergamon Press Ltd.
© 1986 International Federation of Automatic Control

The Modified Gain Extended Kalman Filter and Parameter Identification in Linear Systems*

TAEK L. SONG† and JASON L. SPEYER†

A new nonlinear filter, developed for a special class of nonlinearities, has a universal linearization with respect to the measurement functions and superior convergence and stability characteristics when applied to state and parameter estimation problems in linearized systems.

Key Words—Estimation; filtering; (extended Kalman filters); nonlinear filtering; observers; parameter estimation; state estimation.

Abstract—For a special class of systems, a general formulation and stochastic stability analysis of a new nonlinear filter, called the modified gain extended Kalman filter (MGEKF), is presented. Used as an observer, it is globally exponentially convergent. In the presence of uncertainties a nominal nonrealizable filter algorithm is developed for which global stochastic stability is proven. With respect to this nominal filter algorithm, conditions are obtained such that the effective deviations of the realizable filter are not destabilizing. In an appropriate coordinate frame, the parameter identification problem of a linear system is shown to be a member of this special class. For the example problems, the MGEKF shows superior convergence characteristics without evidence of instability.

1. INTRODUCTION

FOR A SPECIAL class of systems, a general formulation and stochastic analysis of a new nonlinear filter, called the modified gain extended Kalman filter (MGEKF), is presented. The essential idea behind the MGEKF is that the nonlinearities be "modifiable nonlinearities" implying a type of universal linearization. This simple notion, defined in Section 2, is the central idea used in developing the structure of the nonlinear estimator. The analysis for the MGEKF was first given in Song and Speyer (1985) for the case of linear dynamics and modifiable nonlinear measurement function. Here, the general MGEKF is presented for a stochastic system where both the nonlinear measurement functions and nonlinear dynamics are in the class of modifiable functions. The stability analysis presented here is sufficiently different from that of Song and Speyer (1985) to warrant presentation. This generalization is partly motivated by the parameter identification problem in linear systems. It is shown that in an appropriate coordinate frame, the nonlinear dynamics, associated with an augmented state vector which includes the unknown parameters, is a modifiable function.

Recursive identification of system parameters has been widely studied in recent years. Among them are the identification procedures developed by Landau (1976) where his model reference adaptive systems (MRAS) technique is analyzed solely on the bases of deterministic stability considerations by applying hyper-stability concepts to the equivalent feedback representation (EFR) of the algorithm. Ljung (1977) has developed a useful method, called the ordinary differential equation (ODE) method, to analyze the convergence of the recursive identification algorithms in the presence of uncertainty. Ljung (1979) has studied the convergence analysis of the predicted state extended Kalman filter (EKF) applied to the simultaneous identification of the states and parameters of linear systems by using the theory developed in Ljung (1977). Ljung (1979) also suggested a modification to the predicted state EKF algorithm to assure its asymptotic convergence. The analysis in Ljung (1979) requires that the stability of a nonlinear system be tested so as to insure that the estimates of the parameters are stable and convergent to stationary values. These stationary values are obtained by keeping the current estimates of the parameters inside the stability domain defined such that the estimates of the states are exponentially stable. Weiss and Moore (1980) have developed an exponentially convergent estimation algorithm which does not require a stability test by incorporating the Kalman gain calculations inside the system matrix of the state estimates. The convergence characteristics of the estimates are

* Received 18 October, 1983; revised 19 June 1984; revised 16 July 1984. The original version of this paper was not presented at any IFAC Meeting. This paper was recommended for publication in revised form by Associate Editor Y. Sunahara under the direction of Editor P. C. Parks. This research was partially supported by Eglin AFB under contract F08635-82-C-0090, AFOSR under Grant AFOSR-84-0371, and the Joint Services Electronics Program under contract F49620-77-C-0101.

† Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin, Austin, TX 78712, U.S.A.

asymptotically equivalent to the modified algorithm of Ljung (1979). However, the modification in Ljung (1979) and Kalman gain calculations in Weiss and Moore (1980) certainly require more computational effort. Moreover, in those papers, the cause of instability of the algorithm is rather overlooked while concentrating on obtaining stable estimates.

In this paper the MGEKF is applied to the parameter identification of linear systems. In Section 2, the globally convergent observer, called the modified gain extended Kalman observer (MGEKO), is developed for a special system composed of both nonlinear dynamics and nonlinear measurements in an effort to generalize the result of Song and Speyer (1985). The gain algorithm of the MGEKO is the same as that of the Kalman filter. By using Lyapunov's second method, the resulting algorithm is shown to be exponentially convergent. Another exponentially convergent algorithm for the parameter identification problem can be found in Anderson and Johnson (1982). In Section 3, a stability analysis of the MGEKF is studied in the probabilistic Hilbert space L_2 by introducing an exponentially bounded nominal filter called the intermediate MGEKF. Later, sufficient conditions for the MGEKF to be exponentially stable in L_2 are obtained by comparing the estimates of the MGEKF with those of the intermediate MGEKF. Besides the generalization of the results of Song and Speyer (1985), a stability analysis for the parameter identification problem is studied. In Section 5 two examples of the application of the MGEKF to the parameter identification problem are presented. A simple example of Ljung (1979) illustrates that a modification in Ljung (1979) is similar to the MGEKF algorithm, at least for the transient period which is critical to stability of the estimates. Furthermore, the filtered state EKF may have global stability for this example. A convergence analysis of Ljung (1979) is studied for the MGEKF using a simple example and compared with the filtered state EKF analyzed in Westlund and Tysso (1980) and Ursin (1980). However, when the process noise variance is not known exactly, the MGEKF has slightly bigger biases than the filtered state EKF. Finally, the MGEKF is applied to a pole identification problem of a linear time-invariant system excerpted from Saridis (1974) where the EKF is shown to have poor performance. This application also illustrates the robustness of the MGEKF since statistics of the process noise model are mismatched with that of the actual noise. Note that since it is usual to model the parameters as constants, the parameter estimator gains for large time are inversely proportional to time, and therefore, converge to zero. However, our results extend to parameter models which for example are Gauss Markov so that the parameter estimation gains remain finite.

2. THE MODIFIED GAIN EXTENDED KALMAN OBSERVER (MGEKO)

In this section, a globally convergent observer called the modified gain extended Kalman observer (MGEKO) is developed for a class of nonlinear functions (Song and Speyer, 1985). Song and Speyer (1985) have developed the MGEKO for systems with linear dynamics and nonlinear measurement. This is generalized here to the case where both the system dynamics and measurements are nonlinear functions of the system states.

Consider the deterministic discrete-time nonlinear system governed by the following equations

$$x_{i+1} = f_i(x_i) \qquad (2.1)$$

$$z_i^* = h_i(x_i) \qquad (2.2)$$

where $i \in Z_+$ (the non-negative integers), $x_i \in \mathbb{R}^n$ are state variables, and $z_i^* \in \mathbb{R}^q$ are noiseless measurements.

The notion of a modifiable nonlinearity is defined as

Definition 1. A time-varying function $a_i : \mathbb{R}^n \rightarrow \mathbb{R}^p$ is a *modifiable nonlinear system function* if there exists a $p \times n$ time-varying matrix of functions $\mathcal{H}_i : \mathbb{R}^q \times \mathbb{R}^n \rightarrow \mathbb{R}^{p \times n}$ so that for any $x, \bar{x} \in \mathbb{R}^n$ and $i \in Z_+$,

$$a_i(x) - a_i(\bar{x}) = \mathcal{H}_i(z_i^*, \bar{x})(x - \bar{x}) \qquad (2.3)$$

where $z_i^* = h_i(x)$.

Note that $\mathcal{H}(z_i^*, \bar{x}_i)(x_i - \bar{x}_i)$ in (2.3) is a universal linearization with respect to the measurement function $h_i(x_i)$ without any approximation. Notice also that $\mathcal{H}_i(z_i^*, \bar{x}) = \mathcal{H}_i(h_i(x), \bar{x}) \neq \mathcal{H}_i(h_i(\bar{x}), \bar{x})$, where this latter quantity is (if a_i is differentiable) the differential of a_i evaluated at $\bar{x}$, as used in linearization.

Although the class of modifiable functions is small, it contains nonlinear functions used in many practical estimation problems. Two examples are given below which illustrate a modifiable nonlinearity in the dynamics and in the measurement function.

Example 1. Consider the simple linear dynamic system with an unknown coefficient which is analyzed for the noisy case in Section 5.

$$y_{i+1} = \theta_i y_i, \quad \theta_{i+1} = \theta_i, \quad z_i^* = y_i \qquad (2.4)$$

where $x_i^T \doteq [y_i, \theta_i]$. This is easily put into modifiable form as

$$\begin{bmatrix} \theta_i y_i \\ \theta_i \end{bmatrix} - \begin{bmatrix} \bar{\theta}_i \bar{y}_i \\ \bar{\theta}_i \end{bmatrix} = \begin{bmatrix} \theta_i y_i - \bar{\theta}_i y_i + \bar{\theta}_i y_i - \bar{\theta}_i \bar{y}_i \\ \theta_i - \bar{\theta}_i \end{bmatrix}$$
$$= \mathcal{H}_i(z_i^*, \bar{x}_i)(x_i - \bar{x}_i) \qquad (2.5)$$

where

$$\mathcal{H}_i(z_i^*, \bar{x}_i) = \begin{bmatrix} \bar{\theta}_i & z_i^* \\ 0 & 1 \end{bmatrix}, \quad x_i - \bar{x}_i = \begin{bmatrix} y_i - \bar{y}_i \\ \theta_i - \bar{\theta}_i \end{bmatrix}. \quad (2.6)$$

Note $z_i^* = y_i$ is used in (2.6). Note that since no differentiation is assumed, parameters modelled as discrete valued random variables can be included (Marcus and Westwood, 1984).

Example 2. For the noiseless two-dimensional bearings-only measurement problem (Nardon and Aidala, 1981), the system equations are governed by

$$x_{i+1} = Ax_i + Bu_i \quad (2.7)$$

$$z_i^* = \tan^{-1}(Y_i/X_i) \doteq h_i(x_i) \quad (2.8)$$

where A and B are known constant matrices, u_i is a known guidance command, and x_i is the system state consisting of two relative positions (X_i, Y_i), two relative velocities (v_{X_i}, v_{Y_i}), and two target accelerations (a_{TX_i}, a_{TY_i}) such that $x_i = [X, Y, v_X, v_Y, a_{TX}, a_{TY}]_i^T$. Then the nonlinear measurement function $h_i(x_i)$ of (2.8) can be manipulated into the form $h_i(x_i) - (h_i(\bar{x}_i) = \mathcal{H}_i(z_i^*, \bar{x}_i)(x_i - \bar{x}_i)$ where (Speyer and Song, 1981)

$$\mathcal{H}_i(z_i^*, \bar{x}_i) = D_i(\tan^{-1}\alpha_i)/\alpha_i H(z_i^*) \quad (2.9)$$

where $H(z_i^*) = [\sin z_i^*, \ -\cos z_i^*, \ 0, \ 0, \ 0, \ 0]$, $D_i = 1/(\cos z_i^* \bar{X}_i + \sin z_i^* \bar{Y}_i)$, and $\alpha_i = D_i H(z_i^*)\bar{x}_i$. The three-dimensional bearings-only measurement problem is shown to be an approximately modifiable function in Song and Speyer (1985). Other examples of modifiable functions are found in Song (1983).

If the observer for the system of (2.1) and (2.2) is selected in the form of the extended Kalman filter (EKF), then

$$\bar{x}_{i+1} = f_i(\hat{x}_i) \quad (2.10)$$

$$\hat{x}_i = \bar{x}_i + \mathbf{k}_i(z_i^* - h_i(\bar{x}_i)) \quad (2.11)$$

where $\bar{x}_i$ is interpreted as a propagated estimate at time i, $\hat{x}_i$ is interpreted as an updated estimate, and $\mathbf{k}_i$ is some gain sequence that may depend upon past and present data. Suppose $f(\cdot)$ and $h(\cdot)$ of (2.1) and (2.2) are modifiable functions such that

$$f_i(x_i) - f_i(\hat{x}_i) = \mathcal{A}_i(z_i^*, \hat{x}_i)(x_i - \hat{x}_i) \quad (2.12)$$

and

$$h_i(x_i) - h_i(\bar{x}_i) = \mathcal{G}_i(z_i^*, \bar{x}_i)(x_i - \bar{x}_i) \quad (2.13)$$

where $\mathcal{A}_i(z_i^*, \hat{x}_i) \in \mathbb{R}^{n \times n}$, $\mathcal{G}_i(z_i^*, \bar{x}_i) \in \mathbb{R}^{q \times n}$. Then, the actual errors in the estimates of the observer of (2.10) and (2.11) can be written by using (2.1), (2.2), (2.10) and (2.11) as

$$\bar{e}_{i+1} \doteq x_{i+1} - \bar{x}_{i+1} = \mathcal{A}_i(z_i^*, \hat{x}_i)e_i, \quad (2.14)$$

and

$$e_i \doteq x_i - \hat{x}_i = (I - \mathbf{k}_i \mathcal{G}_i(z_i^*, \bar{x}_i))\bar{e}_i \doteq \mathbf{L}_i \bar{e}_i. \quad (2.15)$$

Since (2.14) and (2.15) are written without any approximations and they are in the same form as that of a linear estimation system, a particular gain sequence based upon Kalman filter-type updates is chosen. Specifically,

$$\mathbf{k}_i = m_i \mathcal{G}_i(z_i^*, \bar{x}_i)^T(\mathcal{G}_i(z_i^*, \bar{x}_i) m_i \mathcal{G}_i(z_i^*, \bar{x}_i)^T + \gamma_i)^{-1} \quad (2.16)$$

$$p_i = (I - \mathbf{k}_i \mathcal{G}_i(z_i^*, \bar{x}_i)) m_i (I - \mathbf{k}_i \mathcal{G}_i(z_i^*, \bar{x}_i))^T + \mathbf{k}_i \gamma_i \mathbf{k}_i^T \quad (2.17)$$

$$m_{i+1} \doteq \mathcal{A}_i(z_i^*, \hat{x}_i) p_i \mathcal{A}_i(z_i^*, \hat{x}_i)^T + Q_i. \quad (2.18)$$

Note that if we had the linear estimation problem with dynamics

$$\xi_{i+1} = \mathcal{A}_i(z_i^*, \hat{x}_i)\xi_i + \omega_i \quad (2.19)$$

and measurements

$$y_i = \mathcal{G}_i(z_i^*, \bar{x}_i)\xi_i + v_i \quad (2.20)$$

where the white noise sequences ω_i and v_i have covariances Q_i and γ_i, respectively, then (2.16)–(2.18) would be precise equations for the covariances m_i and p_i of the one-step predicted and filtered estimates of ξ_i, respectively. Here we are assuming that z_i^* and $\bar{x}_i$ are *known*. We will view Q_i and γ_i as design parameters and will call m_i and p_i the "pseudocovariances" of $\bar{e}_i$ and e_i, respectively.

If $(\mathcal{A}_i, \mathcal{G}_i)$ is uniformly observable and $(\mathcal{A}_i, Q_i^{1/2})$ is uniformly controllable, it can be shown that the error dynamics of (2.14) and (2.15) are globally convergent to zero by using the Lyapunov function $V_i(e_i) = e_i^T p_i^{-1} e_i$ in a way that is similar to the approach used by Moore and Anderson (1980) and McGarty (1974). Note that the uniform observability and uniform controllability involve a rank test of the corresponding Gramians. Unfortunately, $(\mathcal{A}_i, \mathcal{G}_i)$ and $(\mathcal{A}_i, Q_i^{1/2})$ for modifiable nonlinear systems are realization dependent such that the tests are not *a priori* checkable.

Remark 1. In the next section white measurement and process noise sequences are added to the measurement (2.2) and dynamics (2.1). If the gain algorithm of (2.16) is used in this stochastic environment, biased estimates are expected, since the gain and the residual of (2.11) are directly correlated (see Song and Speyer, 1985, for additional detail). Therefore, a gain algorithm similar to that of the EKF which ensures that the gain is a function of the *past* measurements only is recommended. However, if the measurement equation is a nonlinear function of the states, the useful relationship between the observability Gramian and the p_i matrix in (2.17) is no longer available. This discussion motivates the assumptions required in the development of the modified gain extended Kalman filter (MGEKF) presented in Section 3.

3. THE MODIFIED GAIN EXTENDED KALMAN FILTER (MGEKF)

In this section we develop the MGEKF and study its stochastic stability. As stated in Remark 1, the gain algorithm of the MGEKF is altered from that of the MGEKO in order to reduce the biases due to direct correlations between the gain and the residual.

Consider the stochastic case where the nonlinear system of (2.1) and (2.2) with additive noise becomes

$$x_i = f_{i-1}(x_{i-1}) + \omega_{i-1} \tag{3.1}$$

$$z_i = h_i(x_i) + v_i \triangleq z_i^* + v_i \tag{3.2}$$

where again $f(\cdot)$ and $h(\cdot)$ are assumed to be modifiable functions and where $\{\omega_i\}$ and $\{v_i\}$ are zero-mean independent noise sequences with finite second moments Q_i and γ_i, respectively. It is further assumed that the ω_js and v_js are independent. Based on Remark 1, the structure of the MGEKF for the above system is similar to that of the MGEKO in Section 2 except that the gain algorithm is altered. Furthermore, since z_i^* in (3.2) is not available because of the measurement noise v_i, z_i^* is replaced by z_i in the gain formulation. If $h_i(\cdot)$ and $f_i(\cdot)$ are modifiable nonlinear functions and $h_i(\cdot)$ is differentiable, then the algorithm of the MGEKF is summarized as

$$\bar{x}_i = f_{i-1}(\hat{x}_{i-1}) \tag{3.3}$$

$$\hat{x}_i = \bar{x}_i + k_i(z_i - h_i(\bar{x}_i)) \tag{3.4}$$

$$m_i = \mathscr{A}_{i-1}(z_{i-1}, \hat{x}_{i-1})p_{i-1}\mathscr{A}_{i-1}(z_{i-1}, \hat{x}_{i-1})^T + Q_{i-1} \tag{3.5}$$

$$k_i = m_i h_{x_i}^T (h_{x_i} m_i h_{x_i}^T + \gamma_i)^{-1} \tag{3.6}$$

$$p_i = (I - k_i\mathscr{G}_i(z_i, \bar{x}_i))m_i(I - k_i\mathscr{G}_i(z_i, \bar{x}_i))^T + k_i\gamma_i k_i^T, \tag{3.7}$$

where $h_{x_i} = \left.\dfrac{\partial h_i(x_i)}{\partial x_i}\right|_{x_i = \bar{x}_i} = \mathscr{G}_i(h_i(\bar{x}_i), \bar{x}_i)$.

Remark 2. Note that if the predicted measurement $h_i(\bar{x}_i)$ is used in (2.16) rather than z_i^*, then the gain in (2.16) reduces to that of (3.6). In this way, the gain of the MGEKF is in the form of the EKF which ensures that the gain is only a function of past measurements. Since the stochastic estimator (3.4) can be rewritten in a modifiable form without any approximation as

$$\hat{x}_i = \bar{x}_i + C_i(x_i - \bar{x}_i) + k_i v_i \tag{3.8}$$

where $C_i \triangleq k_i\mathscr{G}_i(z_i^*, \bar{x}_i)$, the error equation produced from (3.8), (3.1) and (3.3) are *exact*. Although C_i is not implementable, the form of (3.8) is important for our analysis of the behavior of the algorithm. Furthermore, it is critical to the MGEKF performance that $\mathscr{G}_i$ in (3.7) be calculated using z_i. In contrast, the EKF algorithm (Jazwinski, 1970) calculates $\mathscr{G}_i$ using $h_i(\bar{x}_i)$ in (3.7).

3.1. *Stability analysis of the intermediate MGEKF*

In order to facilitate the stochastic stability analysis of the MGEKF, as a first step we employ an unrealizable estimator, which uses z_i^* to calculate the gain k_i. Although this scheme is not implementable for the stochastic case, it forms a nominal to which the implementable filter is compared. This estimator, called "the intermediate MGEKF" for convenience, is given by the following algorithm.

$$\bar{x}_i^* = f_{i-1}(\hat{x}_{i-1}^*) \tag{3.9}$$

$$\hat{x}_i^* = \bar{x}_i^* + k_i^*(z_i - h_i(\bar{x}_i^*)) \tag{3.10}$$

$$m_i^* = \mathscr{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*)p_{i-1}^*\mathscr{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*)^T + Q_{i-1} \tag{3.11}$$

$$k_i^* = m_i^*(h_{x_i}^*)^T((h_{x_i}^*)m_i^*(h_{x_i}^*)^T + \gamma_i)^{-1} \tag{3.12}$$

$$p_i^* = (I - k_i^*\mathscr{G}_i(z_i^*, \bar{x}_i^*))m_i^*(I - k_i^*\mathscr{G}_i(z_i^*, \bar{x}_i^*))^T + k_i^*\gamma_i k_i^{*T} \tag{3.13}$$

where the superscript * is to distinguish the estimates of the intermediate MGEKF from those of the MGEKF. The essential change in forming the intermediate MGEKF is that z_i^* rather than z_i is used to calculate m_i and p_i in (3.5) and (3.7).

We consider the stability of the intermediate MGEKF by using Lyapunov's second method in the probabilistic Hilbert space L_2.

Before proceeding further, the following definition is introduced.

Definition 2 (Tarn and Rasis, 1976). A discrete stochastic process x_i is said to be exponentially bounded in mean square with exponent δ, if there exists constants $0 < \delta < 1$, $K_1 \geq 0$, and $K_2 \geq 0$ such that

$$\|x_i\|^2 \leq K_1 + K_2(1 - \delta)^i. \tag{3.14}$$

$\|x_i\|^2$ in (3.14) is defined in the probabilistic Hilbert space L_2 such that

$$\|x_i\|^2 = \int_{-\infty}^{\infty} x_i^T x_i \pi(x_i) dx_i \tag{3.15}$$

where $\pi(x_i)$ is the probability density function of x_i.

The errors in the estimates of the intermediate MGEKF can be written from (3.1), (3.2), (3.9) and (3.10) as

$$\bar{e}_i^* = x_i - \bar{x}_i^* = \mathscr{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*)e_{i-1}^* + \omega_{i-1} \tag{3.16}$$

and

$$\begin{aligned} \hat{e}_i^* = x_i - \hat{x}_i^* &= (I - K_i^*\mathscr{G}_i(z_i^*, \bar{x}_i^*))\bar{e}_i^* - k_i^* v_i \\ &\triangleq L_i^*\mathscr{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*)e_{i-1}^* - k_i^* v_i + L_i^*\omega_{i-1} \end{aligned} \tag{3.17}$$

where L_i^* is defined as $L_i^* \triangleq I - k_i^*\mathscr{G}_i(z_i^*, x_i^*)$, and where (3.16) is introduced.

Lyapunov functions for e_i^* and $\bar{e}_i^*$ are in the form of

$$V_i(e_i^*) = e_i^{*T} p_i^{*-1} e_i^* \tag{3.18}$$

and

$$\bar{V}_i(\bar{e}_i^*) = \bar{e}_i^{*T} m_i^{*-1} \bar{e}_i^*. \tag{3.19}$$

Before stating Theorem 1, the following assumptions are needed.

Assumption 1: $\mathscr{A}_i(z_i^*, \hat{x}_i^*)$ of (3.16) is uniformly bounded and invertible.
Assumption 2: L_i^* in (3.17) is invertible for all $i \in Z_+$.
Assumption 3: Q_i is uniformly bounded from below such that $Q_i \geq \alpha \cdot I > 0$ for all $i \in Z_+$.
Assumption 4: p_i^{*-1} in (3.18) is bounded from below by a constant matrix $\alpha \cdot I$ for all $i \in Z_+$ such that

$$\|V_i(e_i^*)\| = \|e_i^{*T} p_i^{*-1} e_i^*\| \geq c\|e_i^*\|^2. \tag{3.20}$$

Remark 2. Assumptions 1 and 3 are not terribly restrictive. Note that for the MGEKO described in Section 2, the uniform observability of $(\mathscr{A}_i, \mathscr{G}_i)$ is sufficient to guarantee that L_i defined in (2.15), is invertible and that p_i^{-1} is uniformly bounded from below. The corresponding conditions on L_i^* and p_i^{-1} in Assumptions 2 and 4 are not unreasonable, although there is no such simple a sufficient condition that can be checked.

Theorem 1. The errors in the estimates of the intermediate MGEKF for the system of Equations (3.1) and (3.2) are exponentially bounded in mean square with exponent δ under Assumptions 1, 2, 3 and 4.

Proof. See Appendix 1.

The objective of the following is to show that when the measurement equations of (3.2) is linear in x_i, i.e.

$$z_i = H_i x_i + v_i \tag{3.21}$$

Assumptions 2, 3 and 4 can be relaxed. Instead an observability assumption is used. Relaxing Assumption 3 is important when applying the MGEKF to the parameter identification problem. The algorithm of the intermediate MGEKF for this case can be written as

$$\bar{x}_{i+1}^* = f_i(\hat{x}_i^*) \tag{3.22}$$

$$\hat{x}_i^* = \bar{x}_i^* + k_i^*(z_i - H_i\bar{x}_i^*) \tag{3.23}$$

$$m_i^* = \mathscr{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*)p_{i-1}^*\mathscr{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*)^T + Q_{i-1} \tag{3.24}$$

$$k_i^* = m_i^* H_i^T (H_i m_i^* H_i^T + \gamma_i)^{-1} \tag{3.25}$$

$$p_i^* = (I - k_i^* H_i)m_i^*. \tag{3.26}$$

For the linear measurement case, since $H_i = \mathscr{G}_i = h_{x_i}$, the update formula for p_i^* of the intermediate MGEKF is essentially the same as that of the Kalman filter. Therefore, positive definiteness of Q_i in Theorem 1 can be relaxed to show the global stability of the intermediate MGEKF. With a small modification of the method suggested by Moore and Anderson (1980), the invertibility of $\mathscr{A}_i(z_i^*, \hat{x}_i^*)$ can be relaxed to prove the stability of the intermediate MGEKF. However, here the invertibility of $\mathscr{A}_i(z_i^*, \hat{x}_i^*)$ is kept to develop Theorem 2.

An important aspect of the case of linear measurements is that for $k_i^* = p_i^* H_i^T \gamma_i^{-1}$ the following inequality holds for every $Q_i \geq 0$ and nonsingular $\mathscr{A}_i$ (see the Appendix in Moore and Anderson, 1980, for a more general case)

$$\begin{aligned} p_{i-1}^{*-1} - \mathscr{A}_{i-1}^T L_i^{*T} p_i^{*-1} L_i^* \mathscr{A}_{i-1} \\ \geq \mathscr{A}_{i-1}^T L_i^{*T} H_i^T \gamma_i^{-1} H_i L_i^* \mathscr{A}_{i-1} \end{aligned} \tag{3.27}$$

where $L_i^* = (I - k_i^* H_i)$. Moreover, if $(\mathscr{A}_i, H_i)$ is uniformly observable, p_i^{*-1} is uniformly bounded from below. Therefore, Assumptions 2, 3 and 4 are not needed in the proof of the exponential boundedness of the errors for the intermediate MGEKF.

Note that the Riccati equation for the intermediate MGEKF can also be obtained from minimizing J_N

$$J_N = x_0^T p_0^{*-1} x_0 + \sum_{i=1}^{N} x_i^T H_i^T \gamma_i^{-1} H_i x_i + w_i^T \bar{Q}_i^{-1} w_i \tag{3.28}$$

subject to $x_{i+1} = \mathscr{A}_i x_i + b_i w_i$. The minimum value of J_N is $x_N^T p_N^{*-1} x_N$. Note that $\bar{Q}_i$ of (3.28) is related to

Q_i in the algorithm of the intermediate MGEKF as $Q_i = b_i\bar{Q}_ib_i^T$, and Q_i can be factored as $Q_i = Q_{c_i}Q_{c_i}^T$.

Lemma 1

$Q_c^T p_N^{*-1} Q_c$ is uniformly bounded for all $N \in Z_+$.

Proof. If one chooses $x_c = Q_c y$ where $y \in \Re^n$, then $x_c \in$ Range (W_c), where Range (W_c) denotes the range space of the controllability Gramian W_c. Since Range $(W_c) \supset$ Range (Q_c), x_c is a controllable state. Therefore, from the fact that the minimum cost of (3.28) is finite for any controllable states, $x_c^T p_N^{*-1} x_c = y^T Q_c^T p_N^{*-1} Q_c y < M < \infty$ for all $N \in Z_+$ and for all $y \in \Re^n$.

Before stating Theorem 2 consider the following observability assumption.

Assumption 5: For every k, and some $N > 0$, there exists a $\beta > 0$, such that

$$S_{k+N,k} = \sum_{i=k}^{k+N} \phi_{i,k}^T H_i^T H_i \phi_{i,k} \geq \beta \cdot I > 0 \qquad (3.29)$$

where $\phi_{i,k} = \prod_{j=k}^{i-1} \mathcal{A}_j$. (Note that Assumption 5 implies Assumption 2.)

Theorem 2. If the measurement equation is linear as (3.21), Assumptions 1 and 5 are satisfied, and the w_js and v_js have finite second and fourth moments, then the errors in the estimates of the intermediate MGEKF are exponentially bounded in mean square with exponent δ for positive semidefinite Q_i.

Proof. See Appendix 1.

3.2 *Stability analysis of the MGEKF*

So far the exponential boundedness only of the intermediate MGEKF has been proved. The objective of this section is to obtain sufficient conditions for the MGEKF to be exponentially stable by comparing the estimates of the MGEKF with those of the exponentially stable intermediate MGEKF. Such conditions are found again by using Lyapunov's second method. In this way, the conditions for the deviation from the nominal to belong to the set of nondestabilizing deviations (Safonov, 19[illegible]) are obtained.

The errors in the estimates of the MGEKF can be written from (3.1), (3.2), (3.3) and (3.4) as

$$e_i = x_i - \hat{x}_i = (I - k_i\mathcal{H}_i(z_i^*, \bar{x}_i))\bar{e}_i - k_i v_i \qquad (3.30)$$

where

$$\bar{e}_i = x_i - \bar{x}_i = \mathcal{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1})e_{i-1} + \omega_{i-1}. \qquad (3.31)$$

The only difference between (3.30) and (3.31) of the MGEKF, and (3.16) and (3.17) of the intermediate MGEKF, results from the algorithm for the calculation of the gain filter. That is, the MGEKF uses z_i instead of z_i^* in the gain algorithm. Therefore, $V_{i-1} = \{v_1, v_2, \ldots v_{i-1}\}$ contributes to the difference in the gain calculation and consequently V_{i-1} affects the estimates, $\mathcal{A}_{i-1}$ and $\mathcal{H}_i$. Since $C_i^* = k_i^*\mathcal{H}_i(z_i^*, x_i^*)$ is evaluated by the intermediate MGEKF, then it is convenient to write $C_i = k_i\mathcal{H}_i(z_i^*, x_i)$ of the MGEKF as used in (3.30) as the sum of C_i^* and the perturbation ΔC_i as

$$C_i = C_i^* + \Delta C_i \qquad (3.32)$$

where ΔC_i is dependent on z_i^* and V_{i-1}. Similarly, $\mathcal{A}_{i-1}(z_i^*, \hat{x}_{i-1})$ and k_i of the MGEKF is defined as

$$\mathcal{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}) = \mathcal{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*) + \Delta\mathcal{A}_{i-1}, \qquad (3.33)$$

$$k_i = k_i^* + \Delta k_i. \qquad (3.34)$$

Consider the following sufficiency theorem which states the conditions for the deviations ΔC_i and $\Delta\mathcal{A}_{i-1}$ to belong to the set of nondestabilizing deviations such that the errors in the estimates of the MGEKF are exponentially bounded.

Theorem 3. If Δk_i in (3.34) is bounded and ΔC_i in (3.32) and $\Delta\mathcal{A}_{i-1}$ in (3.33) belong to the set of nondestabilizing deviations such that

$$(I - C_i^* - \Delta C_i)^T p_i^{*-1}(I - C_i^* - \Delta C_i) - m_i^{*-1} \leq 0 \qquad (3.35)$$

$$\begin{aligned}(\mathcal{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*) &+ \Delta\mathcal{A}_{i-1})^T m_i^{*-1}(\mathcal{A}_{i-1}(z_{i-1}^*, \hat{x}_{i-1}^*) \\ &+ \Delta\mathcal{A}_{i-1}) - p_{i-1}^{*-1} < 0 \qquad (3.36)\end{aligned}$$

where p_i^*, m_i^* and p_{i-1}^* are quantities from the intermediate MGEKF, then under the Assumptions 1, 2, 3 and 4 for the system of (3.1) and (3.2), or the Assumptions 1 and 5 for the system of (3.1) and (3.21), the errors in the estimates of the MGEKF are exponentially bounded in mean square with exponent δ.

Proof. See Appendix 1.

Remark 4. Conditions (3.35) and (3.36) can be combined as

$$\begin{aligned}(\mathcal{A}_{i-1} + \Delta\mathcal{A}_{i-1})^T(L_i^* - \Delta C_i)^T p_i^{*-1} \\ (L_i^* - \Delta C_i)(\mathcal{A}_{i-1} + \Delta\mathcal{A}_{i-1}) \\ - p_{i-1}^{*-1} < 0 \qquad (3.37)\end{aligned}$$

where $\mathcal{A}_{i-1}$, $L_i^* = I - k_i^* \mathcal{G}_i(z_i^*, \bar{x}_i^*)$, p_i^* and p_{i-1}^* are quantities from the intermediate MGEKF. If the measurement equation is linear in x_i as (3.21), $\mathcal{G}_i = H_i$ and $\Delta C_i = \Delta k_i H_i$ are used in the above inequality.

Remark 5. Although the global sufficiency conditions (3.35) and (3.36) or (3.37) are uncheckable analytically, they can be used as a guide for engineering evaluation. For example, a local test may be constructed for a given initial state and state estimate. The MGEKF and the intermediate MGEKF algorithms can be run for the same ensemble of measurement and process noise sequences. For each sequence the deviations of Δk_i and ΔC_i can be calculated and the boundedness of Δk_i and the stability of the MGEKF via (3.35) and (3.36) or (3.37) can be assessed.

4. MODIFIABILITY OF THE SYSTEM DYNAMICS OF THE PARAMETER IDENTIFICATION PROBLEM

Consider the following scalar dynamic model

$$z_i^* + \alpha_1 z_{i-1}^* + \ldots + \alpha_n z_{i-n}^* = \beta_1 u_{i-1} + \ldots + \beta_n u_{i-n} \quad (4.1)$$

where n is assumed known and minimal, and u_{i-1} is a known scalar input at time $i-1$. The α_js and β_js are the constant parameters to be identified. The observable canonical form (Chen, 1970) of (4.1) is obtained in the state space Ξ^n as

$$\begin{aligned} y_i &= A y_{i-1} + B u_{i-1} \\ z_i^* &= C y_i \end{aligned} \quad (4.2)$$

where $y_i = [y_1, y_2, \ldots, y_n]_i^T$ and where A, B and C satisfy

$$A = \begin{bmatrix} 0, & 0 & \ldots, & -\alpha_n \\ 1, & 0 & \ldots, & -\alpha_{n-1} \\ & & \vdots & \\ 0, & 0 & \ldots 1, & -\alpha_1 \end{bmatrix}, \quad B = \begin{bmatrix} \beta_n \\ \beta_{n-1} \\ \vdots \\ \beta_1 \end{bmatrix}, \quad C^T = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}. \quad (4.3)$$

If unknown α_js and β_js are augmented to the original state space Ξ^n, the augmented vector x in Ξ^{3n} is

$$x_i^T = [y_1 \ldots y_n, \alpha_n \ldots \alpha_1, \beta_n \ldots \beta_1]_i. \quad (4.4)$$

where x_i satisfies a nonlinear dynamic system expressed in (2.1) but with the addition of a known input u_i.

It is shown below that $f(\cdot)$ corresponding to (4.2), manipulated in the augmented state space Ξ^{3n}, is modifiable. Consider

$$\begin{aligned} A y_{i-1} + B u_{i-1} - \hat{A}_{i-1} \hat{y}_{i-1} - \hat{B}_{i-1} u_{i-1} &= \hat{A}_{i-1}(y_{i-1} - \hat{y}_{i-1}) + (A - \hat{A}_{i-1}) y_{i-1} \\ &\quad + (B - \hat{B}_{i-1}) u_{i-1} \end{aligned} \quad (4.5)$$

$$(A - \hat{A}_{i-1}) y_{i-1} = z_{i-1}^* \begin{bmatrix} -\alpha_n + \hat{\alpha}_n \\ -\alpha_{n-1} + \hat{\alpha}_{n-1} \\ \vdots \\ -\alpha_1 + \hat{\alpha}_1 \end{bmatrix}_{i-1}. \quad (4.6)$$

Therefore, $\mathcal{A}_{i-1}$ in (2.12) for this case can be written as

$$\mathcal{A}_{i-1}(z_{i-1}^*, u_{i-1}, \hat{x}_{i-1}) = \left[\begin{array}{c|c} \hat{A}_{i-1} & -z_{i-1}^* I_n \quad u_{i-1} I_n \\ \hline 0 & I_{2n} \end{array}\right] \quad (4.7)$$

where I_q is the $q \times q$ identity matrix.

5. APPLICATIONS

5.1 *Relation to Ljung's modification and convergence analysis*

Asymptotic behavior of the predicted state EKF as a parameter estimator for linear systems is analyzed in Ljung (1979). In Ljung (1979), the estimate $\bar{\theta}_i$ of the parameter θ is sought for the system

$$y_{i+1} = A(\theta) y_i + \omega_i \quad (5.1)$$

$$z_i = C(\theta) y_i + v_i \quad (5.2)$$

where ω_i and v_i are zero-mean white noise with finite second moments Q_i and γ_i, respectively. The structure of the predicted state EKF, which is an example of the recursive prediction error (RPE) algorithm (Moore and Weiss, 1979), is

$$\bar{x}_{i+1} = f_i(\bar{x}_i) + K_i(z_i - h_i(\bar{x}_i)) \quad (5.3)$$

where the estimate of the augmented state $\bar{x}_i^T = [\bar{y}_i^T, \bar{\theta}_i^T]$. If the derivative of $f_i(\cdot)$ with respect to $\bar{x}_i$ is defined as

$$\frac{\partial f_i(\bar{x}_i)}{\partial \bar{x}_i} = \begin{bmatrix} A(\bar{\theta}_i) & \dfrac{\partial A(\bar{\theta}_i)\bar{y}_i}{\partial \bar{\theta}_i} \\ 0 & I \end{bmatrix} \doteq \begin{bmatrix} A_i & M_i \\ 0 & I \end{bmatrix} \quad (5.4)$$

and the derivative of $h_i(\cdot)$ with respect to $\bar{x}_i$ is defined as

$$\frac{\partial h_i(\bar{x}_i)}{\partial \bar{x}_i} = \begin{bmatrix} C(\bar{\theta}_i) & \dfrac{\partial C(\bar{\theta}_i)\bar{y}_i}{\partial \bar{\theta}_i} \end{bmatrix} \doteq [C_i \vdots D_i], \quad (5.5)$$

and $E\{v_i\omega_j^T\} = 0$ for all i, j, then the gain K_i of the predicted state EKF of (5.3) is obtained from the following algorithm.

$$\begin{aligned} K_i &= [A_iP_{1_i}C_i^T + M_iP_{2_i}^TC_i^T + A_iP_{2_i}D_i^T \\ &\quad + M_iP_{2_i}D_i^T]S_i^{-1} \\ S_i &= C_iP_{1_i}C_i^T + C_iP_{2_i}D_i^T + D_iP_{2_i}^TC_i^T \\ &\quad + D_iP_{3_i}D_i^T + \gamma_i \\ L_i &= [P_{2_i}^TC_i^T + P_{3_i}D_i^T]S_i^{-1} \\ P_{1_{i+1}} &= A_iP_{1_i}A_i^T + A_iP_{2_i}M_i^T + M_iP_{2_i}^TA_i^T \\ &\quad + M_iP_{3_i}M_i^T - K_iS_iK_i^T + Q_i \\ P_{2_{i+1}} &= A_iP_{2_i} + M_iP_{3_i} - K_iS_iL_i^T \\ P_{3_{i+1}} &= P_{3_i} - L_iS_iL_i^T \end{aligned}$$

where the partition of covariance matrix P_i of the state x_i

$$P_i = \begin{bmatrix} P_{1_i} & P_{2_i} \\ P_{2_i}^T & P_{3_i} \end{bmatrix}$$

is used in the algorithm and the initial value of the P_i matrix is $P_0 = \mathrm{diag}\,[P_{1_0}, P_{3_0}]$.

5.1.1 *Ljung's modified EKF*. As shown in Ljung (1979), the predicted state EKF is not asymptotically stable. This fact is analyzed from the stability of the ordinary differential equations obtained by decoupling the estimates and covariance of the state of (5.3) under the assumption that the stationary (or equilibrium) value of $\bar{\theta}$ is obtained. However, the stability domain, defined such that the estimates of the system are exponentially stable, is not explicitly specified in Ljung (1979). In Ljung (1979) the estimator of (5.3) is also modified to be asymptotically stable at the stationary point. This modification includes the change in the gain K_i (see Theorems 7.1 and 8.1, and Corollaries 7.1 and 8.1 of Ljung, 1979) such that

$$\begin{aligned} K_i = [A_iP_{1_i}C_i^T + M_i^*P_{2_i}^TC_i^T + A_iP_{2_i}D_i^T \\ + M_i^*P_{2_i}D_i^T]S_i^{-1} \end{aligned} \qquad (5.6)$$

where

$$M_i^* = M_i + \frac{\partial \bar{K}}{\partial \bar{\theta}}(z_i - C_i\bar{y}_i) \qquad (5.7)$$

where $\bar{K}$ and $\bar{\theta}$ are stationary values of K_i and θ_i, respectively, and $\partial\bar{K}/\partial\bar{\theta}$ is obtained from an approximated algorithm (Equation (7.8) of Ljung, 1979). Since the modified gain K_i of (5.6) is a function of the present measurements, biased estimates are expected during the transient period where P_{2_i} is nonzero (Song and Speyer, 1985; Aidala and Nardon, 1982). For example, in Westlund and Tysso (1980) and Ursin (1980), a simple example from Ljung (1979), i.e.

$$\begin{aligned} y_{i+1} &= \theta_i y_i + \omega_i \\ \theta_{i+1} &= \theta_i \qquad\qquad (5.8) \\ z_i &= y_i + v_i, \end{aligned}$$

is used to show how the filtered state of the EKF (the usual EKF) has asymptotic convergence characteristics by using the same method of Ljung (1979). However, if Ljung's modification (5.6) is applied to the above example, M_i^* in (5.7) becomes

$$M_i^* \cong \bar{y}_i + \frac{P_{1_i}}{P_{1_i} + \gamma_i}(z_i - \bar{y}_i)$$

which can be written as $M_i^* \cong \bar{y}_i + (z_i - \bar{y}_i) = z_i$, if $P_{1_i} > \gamma_i$.

In order to avoid biases in the estimates, the estimator of (5.3) with the modified gain K_i of (5.6) is changed to the form of the filtered state EKF, i.e.

$$\begin{aligned} \bar{x}_i &= f_{i-1}(\hat{x}_{i-1}) \\ \hat{x}_i &= \bar{x}_i + k_i(z_i - H\bar{x}_i), \end{aligned}$$

where $x = [y, \theta]^T$, and $H = [1, 0]$. For the filtered state EKF, Ljung's modification would turn out to be the algorithm of the MGEKF, at least for the transient period which is critical to the stability of the estimates.

Note that $f(\cdot)$ of the above example is modifiable such that $\mathscr{A}_i$ in (2.12) for the system of (5.8) can be written as

$$\mathscr{A}_i(z_i^*, \hat{x}_i) = \begin{bmatrix} \hat{\theta}_i & z_i^* \\ 0 & 1 \end{bmatrix}. \qquad (5.9)$$

In order to calculate the gain of the MGEKF,

$$\mathscr{A}_i(z_i, \hat{x}_i) = \begin{bmatrix} \hat{\theta}_i & z_i \\ 0 & 1 \end{bmatrix} \qquad (5.10)$$

is used, while

$$\frac{\partial f_i}{\partial x_i} = \begin{bmatrix} \hat{\theta}_i & \hat{y}_i \\ 0 & 1 \end{bmatrix} \qquad (5.11)$$

is used in the EKF to calculate the gain. Since the original system of (5.8) is linear, (5.9), (5.10) and (5.11) are similar in form. Since the intermediate MGEKF, which uses $\mathscr{A}_i$ in (5.9) to calculate the gain, is globally stable in the mean square sense by Theorems 1 and 2 of Section 3, the exponential boundedness of the EKF for this example can be obtained by a procedure similar to that used to prove Theorem 3. The result is that the EKF for this example is exponentially bounded if Δk_i, which is the difference between k_i^* of the intermediate MGEKF and k_i of the EKF, is bounded and if the inequality

$$\left(\frac{\partial f_{i-1}}{\partial \hat{x}_{i-1}}\right)^T (I - k_i H)^T p_i^{*-1} (I - k_i H) \left(\frac{\partial f_{i-1}}{\partial \hat{x}_{i-1}}\right) - p_{i-1}^{*-1} < 0 \quad \text{for all } i \in Z_- \tag{5.12}$$

is satisfied where p_i^* and p_{i-1}^* are calculated from the intermediate MGEKF, and $\partial f_{i-1}\, \partial \hat{x}_{i-1}$ and k_i are evaluated from the estimates of the EKF. If the above conditions are always satisfied, then the EKF for this case is globally stable. Otherwise, the regions of satisfaction of (5.12) provide the stability domain for the filtered state EKF. If the conditions of Theorem 3 are satisfied, the errors in the estimates of the MGEKF are exponentially bounded in the mean square and the stationary value of the estimate of parameter θ is obtained in that sense.

5.1.2 *Convergence analysis of the MGEKF.* The following is a convergence analysis of the MGEKF for the scalar system of (5.8). A convergence analysis of the predicted state EKF for the same system was given originally by Ljung (1979), and similar analyses of the filtered state EKF are found in Westlund and Tysso (1980) and Ursin (1980). Satisfaction of the regularity conditions necessary for the proof of convergence is relegated to Appendix I of Ljung (1979). After some manipulation, the MGEKF algorithm which estimates the augmented state x_i of the system (5.8) yields

$$\hat{y}_i = \bar{y}_i + G_i(z_i - \bar{y}_i) \tag{5.13}$$

$$\bar{y}_i = \hat{\theta}_{i-1}\bar{y}_{i-1} \tag{5.14}$$

$$\hat{\theta}_i = \bar{\theta}_i + L_i(z_i - \bar{y}_i) \tag{5.15}$$

$$\bar{\theta}_i = \hat{\theta}_{i-1} \tag{5.16}$$

$$G_i = m_{1_i}(m_{1_i} + \gamma_i)^{-1} \triangleq m_{1_i}s_i^{-1} \tag{5.17}$$

$$L_i = m_{2_i}(m_{1_i} + \gamma_i)^{-1}. \tag{5.18}$$

The matrix m_i is partitioned as

$$m_i = \begin{bmatrix} m_{1_i} & m_{2_i} \\ m_{2_i} & m_{3_i} \end{bmatrix} \tag{5.19}$$

and each partition satisfies

$$\begin{aligned} m_{1_i} = {} & [\hat{\theta}(m_1 - m_1 s^{-1} m_1)\hat{\theta}]_{i-1} \\ & + 2[z(m_2 - m_2 s^{-1} m_1)\hat{\theta}]_{i-1} \\ & + [z(m_3 - m_2 s^{-1} m_2) z]_{i-1} \\ & + Q_{i-1}, m_{1_0} \end{aligned} \tag{5.20}$$

$$\begin{aligned} m_{2_i} = {} & [\hat{\theta}(m_2 - m_1 s^{-1} m_2)]_{i-1} \\ & + [z(m_3 - m_2 s^{-1} m_2)]_{i-1}, \\ m_{2_0} = {} & 0 \end{aligned} \tag{5.21}$$

$$m_{3_i} = [m_3 - m_2 s^{-1} m_2]_{i-1}, m_{3_0}. \tag{5.22}$$

The objective is to demonstrate the local convergence properties of the MGEKF using the linearized differential equation of Ljung (1979) about the stationary point.

It can be shown that m_2 and m_3 tend to be zero as $i \to \infty$ (Ljung, 1979; Moore and Weiss, 1979). Therefore, for $Q = \gamma = 1$ as $i \to \infty$, $m_{1_i} \to \bar{m}_1$ and (5.20) reduces to

$$\bar{m}_1 = \bar{\theta}^2(\bar{m}_1 - \bar{m}_1^2(\bar{m}_1 + 1)^{-1}) + 1. \tag{5.23}$$

This implies that

$$\bar{m}_1 = \frac{\bar{\theta}^2 + \sqrt{\bar{\theta}^4 + 4}}{2} \tag{5.24}$$

where $\bar{\theta}$ is the stationary value of the estimate of the parameter obtained under the assumption that the conditions of Theorem 3 are satisfied. Therefore, $\bar{G}$, which is the limit value of G_i as $i \to \infty$, satisfies using (5.17)

$$\bar{G} = \frac{\bar{m}_1}{\bar{m}_1 + 1}. \tag{5.25}$$

By using (5.22) in (5.21), m_{2_i} satisfies

$$m_{2_{i+1}} = (\hat{\theta}_i - \hat{\theta}_i G_i) m_{2_i} + z_i m_{3_{i+1}}. \tag{5.26}$$

The process $\bar{W}$ satisfying

$$\bar{W}_{i+1} = (\bar{\theta} - \bar{\theta}\bar{G})\bar{W}_i + z_i \tag{5.27}$$

is related to the m_2 process given by (5.26) as $i \to \infty$ with the assumption of constant $\tilde{m}_3$ (Ursin, 1980), i.e.

$$\tilde{m}_{2_i} = [\bar{W}\tilde{m}_3]_i \tag{5.28}$$

where $\tilde{m}_{2_i} = im_{2_i}$ and $\tilde{m}_{3_i} = im_{3_i}$. From (5.15) and (5.16)

$$\hat{\theta}_{i+1} - \hat{\theta}_i = \frac{1}{i}\tilde{L}_i\varepsilon_i \tag{5.29}$$

where $\tilde{L}_i = iL_i$ and $\varepsilon_i = z_i - \hat{\theta}_{i-1}\hat{y}_{i-1}$. From the assumption of ergodicity of the processes and the relation

$$\tilde{L}_i = \tilde{m}_{2_i}(m_{1_i} + 1)^{-1}, \tag{5.30}$$

the RHS of the differential equation associated with θ at the stationary point in Ljung (1979) will be

$$E\{\bar{\tilde{L}}\bar{\varepsilon}\} = E\{\bar{\tilde{m}}_3(\bar{m}_1 + 1)^{-1}\bar{W}\bar{\varepsilon}\}. \tag{5.31}$$

As shown in Ljung (1979), the sign of $E\{\bar{W}\bar{\varepsilon}\}$ decides the stability of $\bar{\theta}$ since $\bar{m}_3(\bar{m}_1+1)^{-1}>0$.

In order to calculate $E\{\bar{W}\bar{\varepsilon}\}$ analytically, transfer functions for $\bar{W}/z$ and $\bar{\varepsilon}/z$ are needed. From (5.27) $\bar{W}/z$ satisfies

$$\frac{\bar{W}}{z}=H_1(q)=\frac{q^{-1}}{1-(\bar{\theta}-\bar{K})q^{-1}} \tag{5.32}$$

where q^{-1} is a one-step delay operator and $\bar{K}=\bar{\theta}\bar{G}$. Note that using the predicted state EKF in Ljung (1979), $\bar{W}/z$ is $\bar{K}H_1(q)^2$ and for the filtered state *EKF* $\bar{W}/z$ is $\bar{G}qH_1(q)^2$. From the relation (limiting equation for (5.13) and (5.14))

$$\hat{y}_i=\bar{\theta}\hat{y}_{i-1}+\bar{G}(z_i-\bar{\theta}\hat{y}_{i-1}),$$

$\hat{y}/z$ is obtained as

$$\frac{\hat{y}}{z}=\frac{\bar{G}}{1-(\bar{\theta}-\bar{K})q^{-1}}. \tag{5.34}$$

Therefore,

$$\frac{\bar{\varepsilon}}{z}=H_2(q)=\frac{1-\bar{\theta}q^{-1}}{1-(\bar{\theta}-\bar{K})q^{-1}}. \tag{5.35}$$

Finally, $E\{\bar{W}\bar{\varepsilon}\}$ is obtained from

$$f_m=E\{\bar{W}\bar{\varepsilon}\}=\frac{1}{2\pi j}\oint_c H_1(q^{-1})H_2(q)\Phi_{zz}(q)\frac{\mathrm{d}q}{q} \tag{5.36}$$

where c is the unit circle in the q domain, and where

$$\Phi_{zz}(q)=1+\lambda/[1-q\theta_0)(1-\theta_0/q)] \tag{5.37}$$

where θ_0 is the actual value of the system parameter and λ is the actual process noise variance, while unit variance is used in the gain calculation of the filter. Note that (5.37) is derived from (5.8) with $\gamma=1$ and the actual process noise variance λ. After evaluation using residue calculus, f_m yields

$$f_m=f_{m_1}+f_{m_2} \tag{5.38}$$

where

$$f_{m_1}=\frac{\lambda(\theta_0-\bar{\theta})\theta_0}{(1-f\theta_0)(\theta_0-f)(1-\theta_0^2)},$$
$$f_{m_2}=\frac{(f-\bar{\theta})[(1-f\theta_0)(f-\theta_0)+\lambda f]}{(1-f^2)(1-f\theta_0)(f-\theta_0)}. \tag{5.39}$$

where $f=\bar{\theta}-\bar{K}=\bar{\theta}/(1+\bar{\theta}^2/2+\sqrt{\bar{\theta}^4/4+1})$. f_m can be compared to the results of the filtered state EKF of Ursin (1980) where f_e of the EKF satisfies $f_e=E\{\bar{W}\bar{\varepsilon}\}=f_{e_1}+f_{e_2}$ where

$$f_{e_1}=\bar{G}f_{m_1}(1-f\theta_0),$$
$$f_{e_2}=\bar{G}f_{m_2}(1-f^2). \tag{5.40}$$

Figures 1 and 2 show the comparison of f_m and f_e for $\theta_0=0.2$, 0.4, 0.6 and 0.8 and for $\lambda=1$ and $\lambda=10$. The results show that there is only one zero crossing for both the filtered state EKF and the MGEKF indicating that there is only one point to which the estimator will converge. Note that for $\lambda=10$ for mismatched process noise variance, the parameter estimate of the MGEKF is only slightly more biased than that of the filtered state EKF. Although the figures show that $|f_m|>|f_e|$ and $|\mathrm{d}f_m/\mathrm{d}\bar{\theta}|>|\mathrm{d}f_e/\mathrm{d}\bar{\theta}|$, this does not indicate that the MGEKF has faster convergence that the EKF, as would be supposed by the analysis of Ursin (1980). The local convergence of the estimators is given by evaluating $E[\bar{L}\bar{\varepsilon}]$ in (5.31) which includes $\bar{m}_3$ and not just by f_m or f_e. Although it can be argued that near the equilibrium point $\bar{m}_3$ is a constant, its value is different for each of the filters and is obtained from Ljung (1979)

$$\bar{m}_3^{-1}\triangleq G(\bar{\theta})=E[\bar{W}^2s^{-1}]$$
$$=\frac{1}{2\pi j}\oint_c H(q^{-1})s^{-1}H(q)\Phi_{zz}(q)\frac{\mathrm{d}q}{q} \tag{5.41}$$

where $H(q)$ is given by H_1 in (5.32) for the MGEKF, by $\bar{G}qH_1(q)^2$ for the filtered state EKF, and by $\bar{\theta}\bar{G}H_1(q)^2$ for the predicted state EKF. All three filters may have similar asymptotic rates of convergence since $\bar{m}_3$ of the MGEKF is less than that of the filtered state EKF, which in turn is less than that of the predicted state EKF. If m_3 is asymptotically the estimation variance, then this indicates that the MGEKF is locally, relatively more efficient than the other EKFs.

5.2 *Pole identification*

Consider the following example of a discrete-time single-input, single-output linear time invariant system excerpted from Saridis (1974)

$$\xi_{i+1}=A\xi_i+dw_i$$
$$z_i=C\xi_i+v_i \tag{5.42}$$

where

$$A=\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\a_1&a_2&a_3&a_4\end{bmatrix},\quad d=\begin{bmatrix}0\\1\\0\\1\end{bmatrix},\quad C^T=\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \tag{5.43}$$

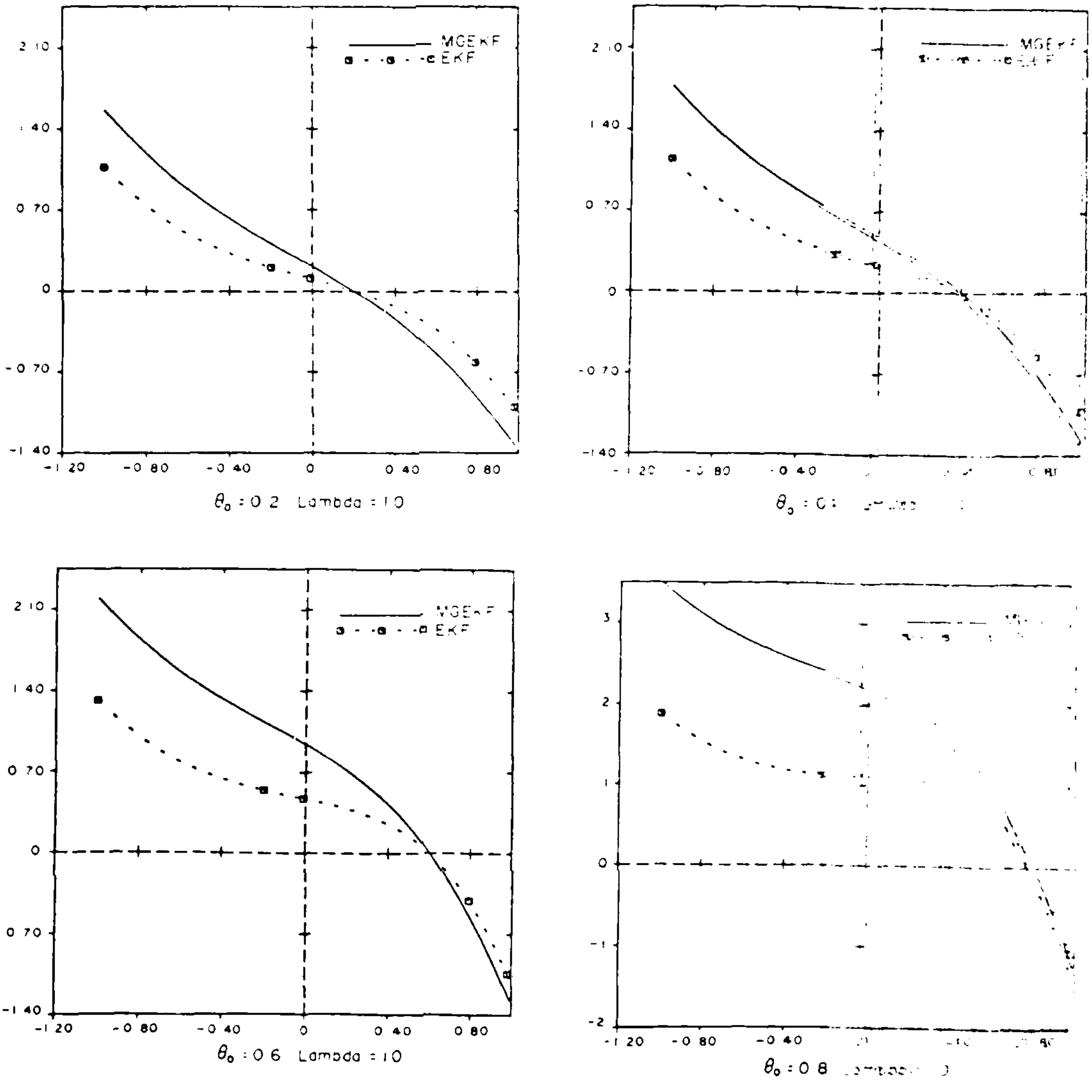


FIG. 1. The functions $f_e(\theta)$ and $f_m(\theta)$ for $\theta_0 = 0.2, 0.4, 0.6$ and 0.8 with

and the noises are zero-mean white Gaussian with finite second moments such as

$$E\{w_i w_j\} = \delta_{ij}$$
$$E\{v_i v_j\} = 0.25\delta_{ij}.$$

According to Saridis (1974), the EKF used to identify the a_js, $j = 1, \ldots 4$ performs poorly even for the case where it is stable. A direct application of the MGEKF to the above system is infeasible, since the nonlinear system dynamics are not modifiable when the unknown parameters are augmented to the original state ξ_t. Therefore, the stochastic system (5.42) is first transformed into the input output transfer function using the Z-transform technique as

$$\frac{z(q)}{w(q)} = \frac{q^2 - a_4 q + (1 - a_3)}{q^4 - a_4 q^3 - a_3 a^2 - a_2 q - a_1}. \quad (5.44)$$

This form is consistent with (4.1). The observable canonical form (Chen, 1970) is realized as

$$\begin{aligned} x_{t+1} &= \tilde{A}x_t + \tilde{d}w_t \\ z_t &= \tilde{c}^T x_t + v_t \end{aligned} \quad (5.45)$$

where

$$\tilde{A} = \begin{bmatrix} 0 & 0 & 0 & a_1 \\ 1 & 0 & 0 & a_2 \\ 0 & 1 & 0 & a_3 \\ 0 & 0 & 1 & a_4 \end{bmatrix}, \quad \tilde{d} = \begin{bmatrix} 1 - a_3 \\ -a_4 \\ 1 \\ 0 \end{bmatrix}, \quad \tilde{c}^T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad (5.46)$$

This is the form suggested in (4.2) and (4.3). If unknown parameters a_1, a_2, a_3 and a_4 are augmented to the original states as x_5, x_6, x_7 and x_8, respectively, then the augmented nonlinear system

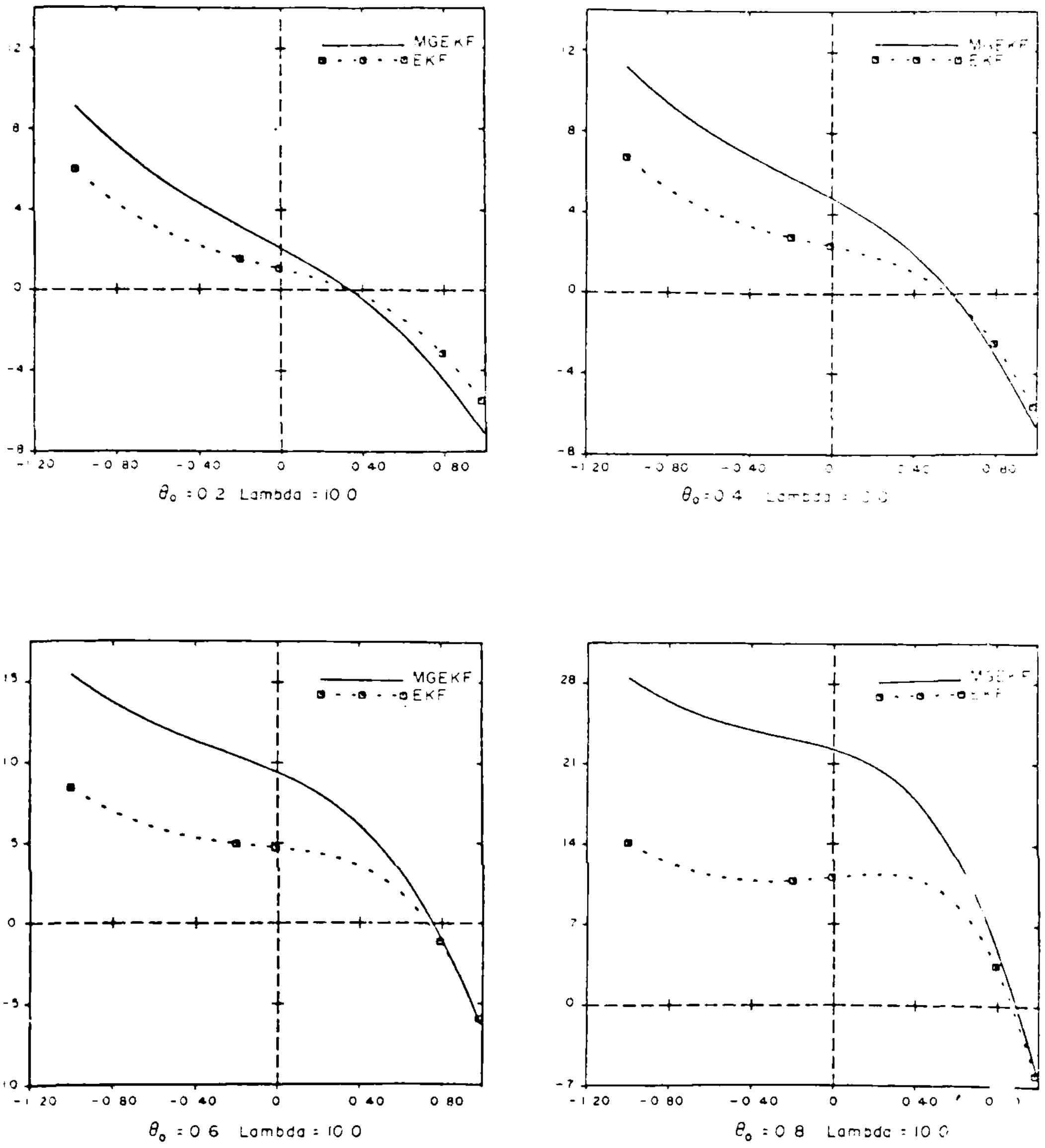


FIG. 2. The functions $f_e(\vartheta)$ and $f_m(\vartheta)$ for $\theta_0 = 0.2, 0.4, 0.6$ and 0.8 with $\lambda = 10$.

dynamics are written as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix}_{i+1} = \begin{bmatrix} x_5 x_4 \\ x_1 + x_6 x_4 \\ x_2 + x_7 x_4 \\ x_3 + x_8 x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix}_i + \begin{bmatrix} (1 - x_7) \\ -x_8 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_i w_i$$

$$\triangleq f_i(x_i) + W_i(x_i, w_i). \qquad (5.47)$$

Following the arguments of Section 4, the measurement $z_i = x_{4_i} + v_i$, and $f(\cdot)$ of (5.47) are modifiable. Note that the above nonlinear system is corrupted by state dependent noise with second moment

$$Q_i = E\{W_i(x_i, w_i) W_i^T(x_i, w_i)\}. \qquad (5.48)$$

Since $(1 - x_{7_i})$ and x_{8_i} in W_i of (5.47) are actually unknown values for this problem, Q_i in (5.48) is first assumed to be an arbitrary constant matrix when the MGEKF is applied, and later, an adaptive scheme, used to calculate Q_i, is implemented in the simulation. The stability analysis in Section 3 is still valid even though the noise model is different from the actual noise as long as the actual noise and noise model are zero-mean processes with finite second moments. As given in Section 4, $Z_i(z_i^*, \hat{x}_i)$, obtained from $f_i(x_i) - f_i(\hat{x}_i)$ is

$$\mathscr{A}_i(z_i^*, \hat{x}_i) = \begin{bmatrix} 0 & 0 & 0 & \hat{x}_{5_i} & x_{4_i} & 0 & 0 & 0 \\ 1 & 0 & 0 & \hat{x}_{6_i} & 0 & x_{4_i} & 0 & 0 \\ 0 & 1 & 0 & \hat{x}_{7_i} & 0 & 0 & x_{4_i} & 0 \\ 0 & 0 & 1 & \hat{x}_{8_i} & 0 & 0 & 0 & x_{4_i} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.49}$$

where $z_i^* \triangleq x_{4_i}$. Note that $\mathscr{A}_i(z_i, \hat{x}_i)$, which is obtained from replacing x_{4_i} in (5.49) by z_i, is used to calculate the gain of the MGEKF, while $\partial f_i / \partial \hat{x}_i$, which is in the same form as (5.49) but with x_{4_i} replaced by $\hat{x}_{4_i}$, is used to calculate the gain of the EKF.

For simulation, the actual values for a_is are selected as $[a_1, a_2, a_3, a_4] = [-0.66, 0.78, -0.18, 1.0]$ as given in Saridis (1974). The matrix Q_i in (5.48) is selected for the matched and mismatched case. An 8×8 Q_i matrix for matched process noise statistics, denoted Q_m, satisfies

$$Q_m = \left[\begin{array}{ccc|c} (1.18)^2, & -1.18, & 1.18 & \\ -1.18\ , & 1\ , & -1 & 0 \\ 1.18\ , & -1\ , & 1 & \\ \hline & 0 & & 0 \end{array}\right]. \tag{5.50}$$

For the mismatched case, the first two diagonal elements of Q_i are increased by 20, 40, 60, 80 and 100%. A series of simulations is executed with matched and mismatched process noise statistics. Figure 3 shows the results of 10 runs of Monte Carlo

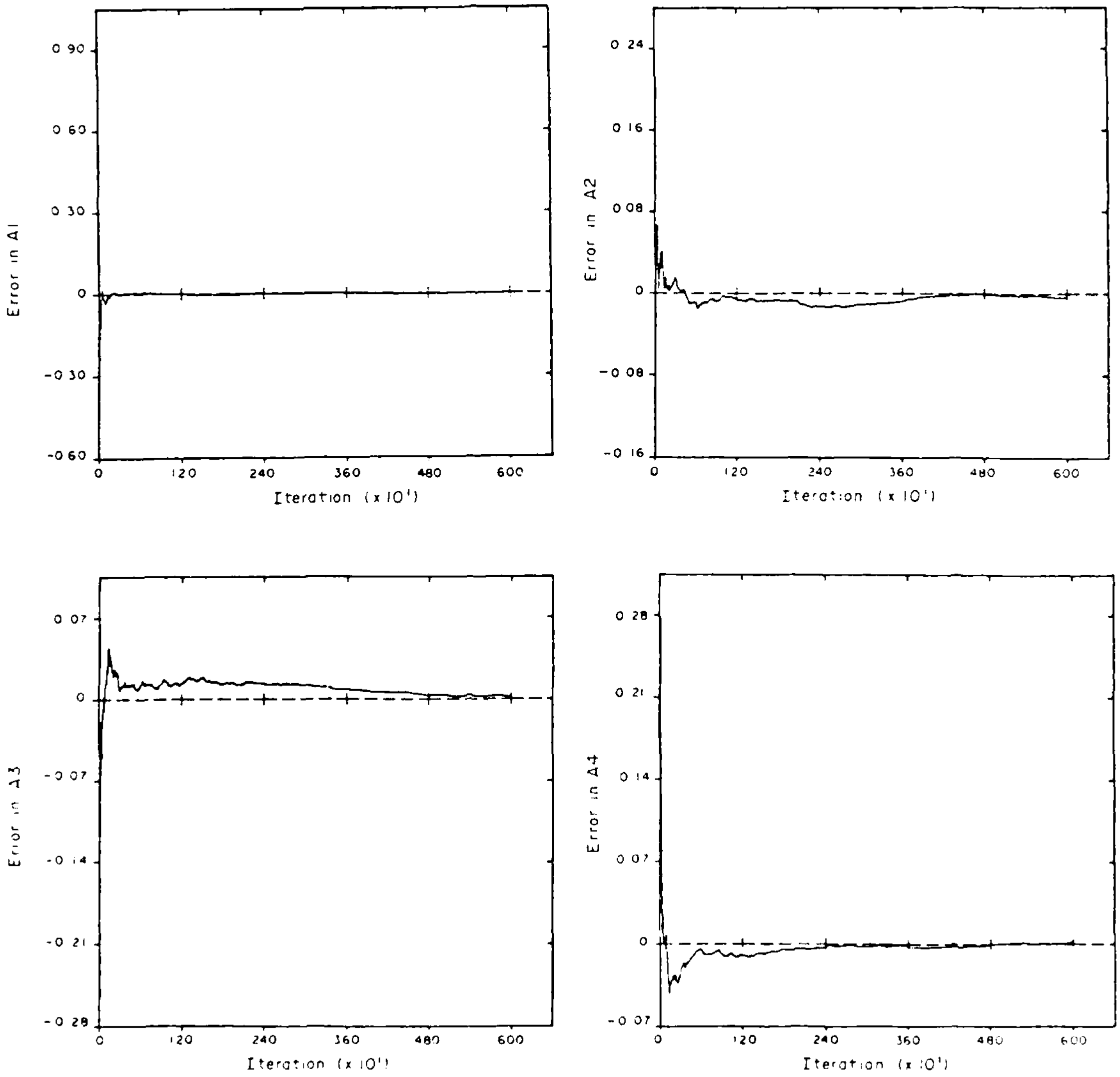


FIG. 3. The errors in the estimates of the parameters in the case of matched process noise variance Q_m.

simulation of the errors in the parameter estimates with the matched process noise statistics, while Fig. 4 shows those of the 100% mismatched cases where

$$Q_{t_{100}} = \left[\begin{array}{ccc:c} (1.669)^2, & -2.36, & 1.669 & \\ -2.36, & 2, & -1.414 & 0 \\ 1.669, & -1.414, & 1 & \\ \hdashline & 0 & & 0 \end{array}\right] \tag{5.51}$$

is used. For all of the simulations, the initial value of true state is $x_0^T = [0, 0, 0, 0, -0.66, 0.78, -0.18, 1.0]$, the initial value of the *a priori* estimation of the state is $\hat{x}_0^T = [5, 5, 5, 5, -1.32, 1.56, -0.36, 2]$, and the initial value of p_i is $p_0 = 10I_8$. Finally, for the purpose of comparison with the results of Saridis (1974), Fig. 5 shows the convergence rates of the MGEKF with Q_{t_0}, $Q_{t_{100}}$ and the adaptive calculations of Q_t. For adaptive calculations of Q_t, the following Q_t is used

$$Q_t = \left[\begin{array}{ccc:c} (1-\hat{x}_7)^2, & -\hat{x}_8(1-\hat{x}_7), & (1-\hat{x}_7) & \\ -\hat{x}_8(1-\hat{x}_7), & \hat{x}_8^2, & -\hat{x}_8 & 0 \\ (1-\hat{x}_7), & -\hat{x}_8, & 1 & \\ \hdashline & 0 & & 0 \end{array}\right] \tag{5.52}$$

The results here show a remarkable improvement over those reported in Saridis (1974) for the EKF using the stochastic system (5.42). The average normalized parameter error was reduced by three orders of magnitude of the EKF of Saridis (1974) and, in fact, the result of the MGEKF is equivalent to the best parameter identification scheme reported here.

With the same arguments as those given in

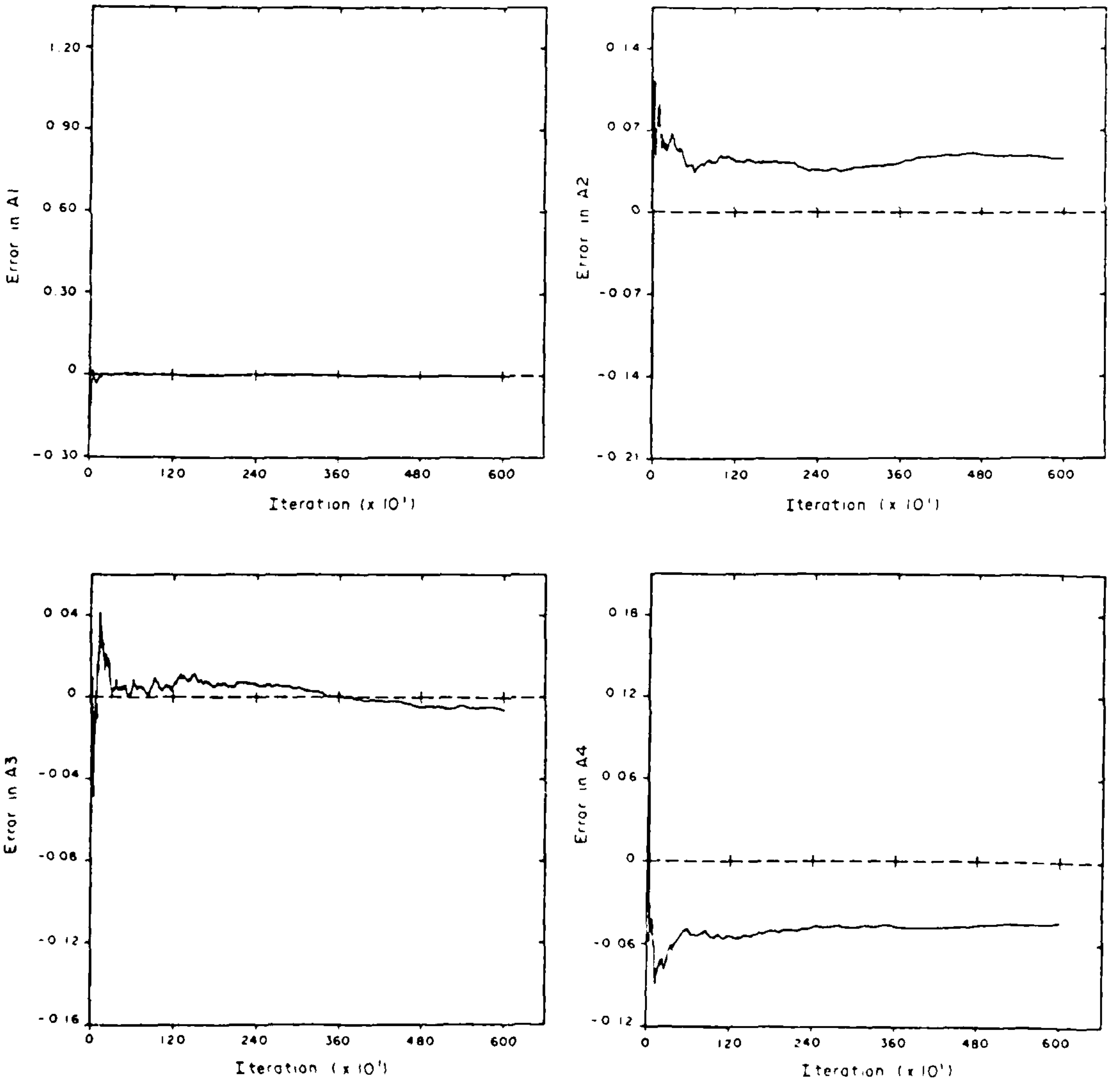


FIG. 4. The errors in the estimates of the parameters in the case of mismatched process noise variance $Q_{t_{100}}$.

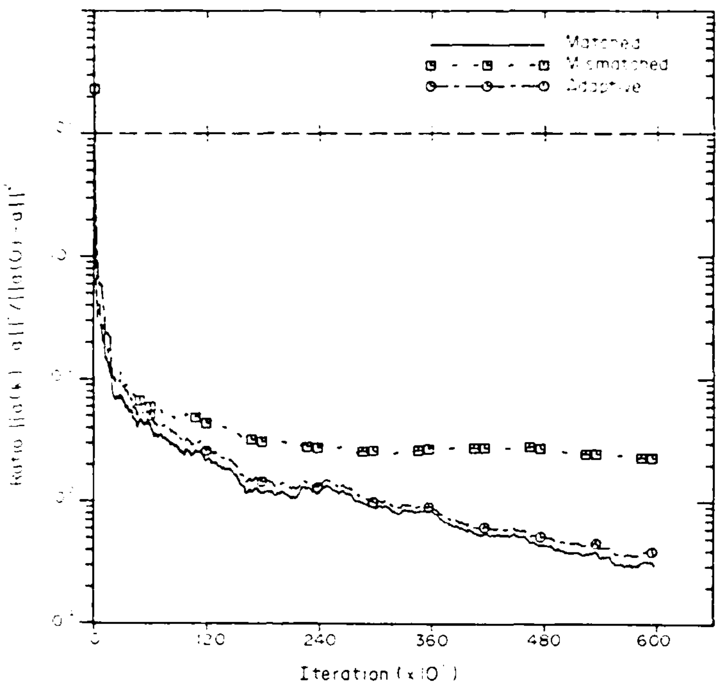


FIG. 5. Convergence rates of the MGEKF with $Q_{(0)}$, $Q_{(100)}$, and adaptive calculations of Q_t.

Section 5.1, the EKF applied to observable canonical form (5.45) may have global stability since the difference between $\mathscr{A}_t(z_t^*, \hat{x}_t)$ of the intermediate MGEKF and $\partial f_t/\partial \hat{x}_t$ of the EKF is small if $x_{4_t} - \hat{x}_{4_t}$ for the EKF is small such that (5.12) in Section 5.1 holds.

6. CONCLUSIONS

An exponentially convergent observer called the MGEKO is derived for the problem where both the nonlinearities in the dynamics and measurement are assumed modifiable. The exponential convergence of the MGEKF is studied in the probabilistic Hilbert space L_2, by introducing the exponentially bounded nominal filter called the intermediate MGEKF, and sufficient conditions for the MGEKF to be globally stable are obtained from the analysis. The intermediate MGEKF can also serve as a nominal filter for the stability analysis of the filtered state EKF used for the modifiable systems. These results generalize the work reported in Song and Speyer (1985) for the case where only the measurement function was nonlinear and modifiable. The stochastic stability analysis for the continuous time MGEKF is found in Song (1983). These generalized results are now applicable to the parameter identification problem where the measurement is linear and the dynamics are nonlinear and modifiable, if the proper coordinate frame is chosen.

Ljung's convergence analysis is applied to a scalar parameter identification problem. The results show that the MGEKF has similar convergence properties as the filtered state EKF which differs markedly from that of the predicted state EKF (Ljung, 1979). The MGEKF is applied to a pole identification problem of a fourth order system where the previous reported results for the performance of the EKF are poor. By choosing the observable canonical form of the state, the resulting system dynamics are modifiable such that the MGEKF is readily applied. Monte Carlo simulation indicates that for the case of matched noise variance and for the case of the adaptive calculations of the noise variance, the MGEKF has excellent convergence characteristics. In the case of mismatched noise variance without the adaptive feature, the MGEKF has decent mean square error performance characteristics.

REFERENCES

Aidala, V. J. and S. C. Nardone (1982). Biased estimation properties of the pseudolinear tracking filter. *IEEE Trans. Aerospace and Electronic Syst.*, **AES-18** (4).

Anderson, B. D. O. and C. R. Johnson (1982). Exponential convergence of adaptive identification and control algorithms. *Automatica*, **18** (1).

Chen, C. T. (1970). *Introduction to Linear System Theory*. Holt, Rinehart & Winston, New York.

Landau, I. D. (1976). Unbiased recursive identification using model reference adaptive techniques. *IEEE Trans. Aut. Control*, **AC-21** (2).

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Aut. Control*, **AC-22** (4).

Ljung, L. (1979). Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems. *IEEE Trans. Aut. Control*, **AC-24** (1).

Marcus, S. I. and E. K. Westwood (1984). On asymptotic approximations for some nonlinear filtering problems. *Proc. IFAC Triennial Congress*, Vol. VII, pp. 36–41. Budapest, Hungary.

McGarty, T. P. (1974). *Stochastic Systems and State Estimation.* Wiley, New York.

Moore, J. B. and B. D. O. Anderson (1980). Coping with singular transition matrices in estimation and control stability theory. *Int. J. Control,* **31** (3).

Moore, J. B. and H. Weiss (1979). Recursive prediction error methods for adaptive estimation. *IEEE Trans. Syst., Man, Cybernetics,* **SMC-9** (4).

Nardon, S. C. and J. J. Aidala (1981). Observability criteria for bearings-only target motion analysis. *IEEE Trans. Aerospace and Electronic Syst.,* **AES-17** (2).

Safonov, M. G. (1980). *Stability and Robustness of Multivariable Feedback Systems.* MIT Press, Cambridge, Mass.

Saridis, G. N. (1974). Comparison of six on-line identification algorithms. *Automatica,* **10**.

Song, T. L. (1983). A stochastic analysis of a modified gain extended Kalman filter. Ph.D. dissertation, University of Texas at Austin.

Song, T. L. and J. L. Speyer (1985). A stochastic analysis of a modified gain extended Kalman filter with applications to estimation with bearings only measurements. *IEEE Trans. Aut. Control,* **AC-30** (10).

Speyer, J. L. and Song, T. L. (1981). A comparison between pseudomeasurement and extended Kalman observers. *Proc. 20th IEEE Conf. Decision and Control,* San Diego, California.

Tarn, T. J. and Y. Rasis (1976). Observers for nonlinear stochastic systems. *IEEE Trans. Aut. Control,* **AC-21** (4).

Ursin, B. (1980). Asymptotic convergence properties of the extended Kalman filter using filtered state estimates. *IEEE Trans. Aut. Control,* **AC-25** (6).

Weiss, H. and J. B. Moore (1980). Recursive prediction error algorithms without a stability test. *Automatica,* **16**.

Westlund, T. and A. Tysso (1980). Remarks on asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems. *IEEE Trans. Aut. Control,* **AC-25** (5).

APPENDIX 1: PROOFS OF THEOREMS 1, 2 AND 3

1. *Proof of Theorem* 1

Take the conditional expectation over $V_i(e_i^*) - \bar{V}_i(\bar{e}_i^*)$ given

$\bar{Y}_i^* = \{e_0^*, \bar{e}_1^*, e_1^*, \ldots e_{i-1}^*, \bar{e}_i^*\}$ as

$$E_{\bar{Y}_i}^*\{V_i(e_i^*) - \bar{V}_i(\bar{e}_i^*)\} = \bar{e}_i^* E_{\bar{Y}_i}^*\{L_i^{*T} p_i^{*-1} L_i^* - m_i^{*-1}\}e_i^* + \mathrm{tr}\,[\mathrm{E}_{\bar{Y}_i}^*\{k_i^{*T} p_i^{*-1} k_i^* \gamma_i\}] \qquad (A.1)$$

$E_{\bar{Y}}^*\{\cdot\}$ is a conditional expectation operator given $\bar{Y}_i^*$, L_i^* is defined in (3.17), and tr is a trace operator for matrices. The derivation of (A.1) uses the facts that z_i^* is not a function of ω_i, $\bar{x}_i^*$ is a function of the past measurements such that L_i^* in (3.17) and p_i^* are independent of v_i, and v_i is a zero-mean independent noise process with finite second moment γ_i. Note that the term inside the first $E_{\bar{Y}_i}^*$ operation of the RHS of (A.1) satisfies the following equation

$$L_i^{*T} p_i^{*-1} L_i^* - m_i^{*-1} = -s_i^{*T}(s_i^* m_i^* s_i^{*T} + \gamma_i)^{-1} s_i^* \le 0 \qquad (A.2)$$

where s_i^* of the intermediate MGEKF satisfies $s_i^* = \gamma_i k_i^{*T} L_i^{*-T} m_i^{*-1}$. The last inequality of (A.2) comes from the fact that $s_i^* \in \mathscr{R}^{q \times n}$ is at most of rank q. Equation (A.2) implies that there exists $0 \le \rho_{1_i} < 1$ such that

$$\bar{e}_i^{*T} L_i^{*T} p_i^{*-1} L_i^* \bar{e}_i^* = \rho_{1_i} e_i^{*T} m_i^{*-1} \bar{e}_i^* = \rho_{1_i} \bar{V}_i(\bar{e}_i^*). \qquad (A.3)$$

Therefore, (A.1) becomes

$$E_{\bar{Y}_i}^*\{V_i(e_i^*) - \bar{V}_i(\bar{e}_i^*)\} = K_{1_i} - (1 - \rho_{1_i}) E_{\bar{Y}_i}^*\{\bar{V}_i(\bar{e}_i^*)\} \qquad (A.4)$$

where K_1 is defined as $K_{1_i} = \mathrm{tr}\,\{\mathrm{E}_{\bar{Y}_i}^*\{k_i^{*T} p_i^{*-1} k_i^* \gamma_i\}\}$, and $0 \le K_{1_i} < M < \infty$. Similarly, take the conditional expectation over $V_i(e_i^*) - V_{i-1}(e_{i-1}^*)$ given $Y_{i-1}^* = \{e_0^*, e_1^*, e_2^*, \ldots e_{i-1}^*\}$, as

$$E_{Y_{i-1}^*}\{V_i(e_i^*) - V_{i-1}(e_{i-1}^*)\} = e_{i-1}^{*T} E_{Y_{i-1}^*}\{\mathscr{A}_{i-1}^T m_i^{*-1} \mathscr{A}_{i-1} - p_{i-1}^{*-1}\} e_{i-1}^* + \mathrm{tr}\,\{E_{Y_{i-1}^*}\{m_i^{*-1} Q_{i-1}\}\} \qquad (A.5)$$

where use is made of the facts that z_{i-1}^* is not a function of ω_{i-1}, $\bar{x}_{i-1}^*$ is a function of the past measurements such that $\mathscr{A}_{i-1} = \mathscr{A}_{i-1}(z_{i-1}^*, \bar{x}_{i-1}^*)$ and m_i^* are independent of ω_{i-1}, and ω_{i-1} is a zero-mean, independent noise process with finite second moment Q_{i-1}. The term inside the first $E_{Y_{i-1}^*}$ of the RHS of (A.5) satisfies

$$\mathscr{A}_{i-1}^T m_i^{*-1} \mathscr{A}_{i-1} - p_{i-1}^{*-1} = -p_{i-1}^{*-1} \mathscr{A}_{i-1}^{-1}(Q_{i-1}^{-1} + \mathscr{A}_{i-1}^{-T} p_{i-1}^{*-1} \mathscr{A}_{i-1}^{-1})^{-1} \mathscr{A}_{i-1}^{-T} p_{i-1}^{*-1} < 0 \qquad (A.6)$$

Equation (A.6) implies that there exists $0 < \beta_1 < \rho_{2_i} < 1$ such that

$$e_{i-1}^{*T} \mathscr{A}_{i-1}^T m_i^{*-1} \mathscr{A}_{i-1} e_{i-1}^* = \rho_{2_i} e_{i-1}^{*T} p_{i-1}^{*-1} e_{i-1}^*. \qquad (A.7)$$

The existence of β_1 is assured by Assumptions 3 and 4. Therefore, (A.5) becomes

$$E_{Y_{i-1}^*}\{\bar{V}_i(\bar{e}_i^*) - V_{i-1}(e_{i-1}^*)\} = K_{2_i} - (1 - \rho_{2_i}) E_{Y_{i-1}^*}\{V_{i-1}(e_{i-1}^*)\} \qquad (A.8)$$

where $K_{2_i} = \mathrm{tr}\,\{\mathrm{E}_{Y_{i-1}^*}\{m_i^{*-1} Q_{i-1}\}\}$, and $0 < K_{2_i} < N < \infty$. Now, take the conditional expectation over (A.4) for given Y_{i-1}^*, and use the nesting property of the conditional expectation, then

$$\begin{aligned} E_{Y_{i-1}^*}\{V_i(e_i^*) - \bar{V}_i(\bar{e}_i^*)\} &= E_{Y_{i-1}^*}\{E_{\bar{Y}_i}\{V_i(e_i^*) - \bar{V}_i(e_i^*)\}\} \\ &= K_{1_i} - (1 - \rho_{1_i}) E_{Y_{i-1}^*}\{(\mathscr{A}_{i-1} e_{i-1}^* + \omega_{i-1})^T m_i^{*-1} \\ &\quad (\mathscr{A}_{i-1} e_{i-1}^* + \omega_{i-1})\} \\ &= K_{1_i} - (1 - \rho_{1_i}) e_{i-1}^* E_{Y_{i-1}^*}\{\mathscr{A}_{i-1}^T m_i^{*-1} \mathscr{A}_{i-1}\} e_{i-1}^* \\ &\quad - (1 - \rho_{1_i}) K_{2_i} \end{aligned} \qquad (A.9)$$

where the fact that ω_{i-1} is independent of $\mathscr{A}_{i-1}$ and m_i^* is used. From (A.9) and (A.8)

$$\begin{aligned} &E_{Y_{i-1}^*}\{V_i(e_i^*) - V_{i-1}(e_{i-1}^*)\} \\ &\quad = K_{1_i} - (1 - \rho_{2_i}) E_{Y_{i-1}^*}\{V_{i-1}(e_{i-1}^*)\} \\ &\qquad - (1 - \rho_{1_i}) e_{i-1}^{*T} E_{Y_{i-1}^*}\{\mathscr{A}_{i-1}^T m_i^{*-1} \mathscr{A}_{i-1}\} e_{i-1}^* \\ &\qquad + \rho_{1_i} K_{2_i} \le \mathbf{K}_1 - \delta_i E_{Y_{i-1}^*}\{V_{i-1}(e_{i-1}^*)\} \end{aligned} \qquad (A.10)$$

where $0 \le \sup_{j \in Z_+} (K_{1_j} + \rho_{1_j} K_{2_j}) \le \mathbf{K}_1 < \infty$, and $\delta_i = 1 - \rho_{2_i}$ such that $0 < \beta_2 \le \delta_i < 1$. The boundedness of $\mathbf{K}_1$ is obtained from Assumptions 3 and 4, and the gain k_i^* algorithm of the intermediate MGEKF. Note that Assumption 3 implies m_i^{*-1} is uniformly bounded from above.

By applying the nesting property of the conditional expectation to (A.10), one can obtain

$$\begin{aligned} E_{Y_{0,i}^*}\{V_i(e_i^*)\} &= E_{Y_{0,i}^*}\{E_{Y_{i-1}^*}\{V_i(e_i^*)\}\} \\ &\le \mathbf{K}_1 + (1 - \delta_i) E_{Y_{0,i}^*}\{E_{Y_{i-1}^*}\{V_{i-1}(e_{i-1}^*)\}\} \\ &= \mathbf{K}_1 + (1 - \delta_i) E_{Y_{0,i}^*}\{V_{i-1}(e_{i-1}^*)\}. \end{aligned} \qquad (A.11)$$

Define δ as $\delta = \inf_{j \in Z_+} \{\delta_j\}$, then δ is uniformly greater than 0. Applying (A.11) recursively results in

$$E_{Y_{0,i}^*}\{V_i(e_i^*)\} \le \mathbf{K}_1 \sum_{j=0}^{i-1} (1 - \delta)^j + (1 - \delta)^i E_{Y_{0,i}^*}\{V_0(e_0^*)\}. \qquad (A.12)$$

Use Assumption 4 and take an unconditional expectation over Y_0^* in (A.12). Then,

$$\|e_i^*\|^2 \le K_1 + K_2(1 - \delta)^i \qquad (A.13)$$

where $K_1 < \mathbf{K}_1 \sum_{j=0}^{i} (1-\delta)^j c = \mathbf{K}_1 c/\delta$ and $K_2 = E\{V_0(e_0^*)\} c$. Therefore, the claim of Theorem 1 is proved. (q.e.d.)

2. *Proof of Theorem 2*

Define a Lyapunov function as $V_k(e_k^*) = e_k^{*T} p_k^{*-1} e_k^*$. After some manipulations, using (3.27) for each interval in $[k, k+N]$, it can be shown that the conditional expectation of $V_{k+N}(e_{k+N}^*) - V_k(e_k^*)$ given $Y_k^* = \{e_0^*, e_1^*, \ldots, e_k^*\}$ satisfies

$$
\begin{aligned}
E_{Y_k^*}\{V_{k+N} - V_k\} \le{}& -e_k^{*T} E_{Y_k^*}\left\{\sum_{j=k}^{k+N} \psi_{j,k}^T \tilde{H}_j \psi_{j,k}\right\} e_k^* \\
&+ \sum_{j=k+1}^{k+N} \mathrm{tr}\,[k_j^{*T} p_j^{*-1} k_j^* \gamma_j] + \mathrm{tr}\,[L_j^{*T} p_j^{*-1} L_j^* Q_{j-1}] \\
&- 2e_k^{*T} \sum_{j=k+1}^{k+N} E_{Y_k^*}\left\{\psi_{j,k}^T \tilde{H}_j \sum_{i=k+1}^{j} \psi_{j,i} T_i\right\} \\
&+ \sum_{j=k+1}^{k+N} E_{Y_k^*}\left\{\sum_{i=k+1}^{j} T_i^T \psi_{j,i}^T \tilde{H}_j \psi_{j,i} T_i\right\}
\end{aligned}
\qquad (A.14)
$$

where $\tilde{H}_j = H_j^T \gamma_j^{-1} H_j$, $T_i = k_i^* v_i - L_i^* \omega_{i-1}$ and $\psi_{j,i} = \prod_{l=i}^{j-1} L_{l+1}^* \Lambda_l$. Furthermore, $e_k^* = \psi_{k,j} e_j^* - \prod_{i=j+1}^{k} \psi_{k,i} T_i$ and the fact that the v_is and ω_is are zero-mean, independent noise sequences are used in the derivation. Note that in the last two terms of (A.14), there is correlation between $\psi_{j,i}$ and the random noises. If Schwarz's inequality is applied to each element of the vector $E_{Y_k^*}\{w_j\}$ where $w_j = \psi_{j,k}^T \tilde{H}_j \sum_{i=k+1}^{j} \psi_{j,i} T_i$, it can be shown that the $E_{Y_k^*}\{w_j\}$ is bounded. Since Assumption 5 is satisfied (observability condition), p_i^* is bounded from above such that k_i^* is bounded. Note that the v_is and ω_is are zero-mean independent white noises with finite second moments. Therefore, there exists a constant vector B_k such that $\sum_{j=k+1}^{k+N} E_{Y_k^*}\{w_j\} = B_k$. Similarly, if Schwarz's inequality and the hypothesis of the boundedness of fourth moments of the noises are applied to the last term of (A.14), it is possible to find a constant K_0 which bounds the last term of (A.14). Note also that since Assumption 5 is satisfied, the argument of the expectation operator of the first term of the RHS of (A.14) is strictly uniformly positive definite. The terms involving the tr operator in (A.14) can be written as

$$
\mathrm{tr}\,[\gamma_j^{-1} H_j p_j^* H_j^T] + \mathrm{tr}\,[p_j^{*-1} Q_{j-1} - 2H_j^T \gamma_j^{-1} H_j Q_{j-1} + H_j^T \gamma_j^{-1} H_j p_j^{*T} \gamma_j^{-1} H_j Q_{j-1}]. \qquad (A.15)
$$

Then, from Lemma 1 in Section 3, $\mathrm{tr}(p_j^{*-1} Q_{j-1})$ of (A.15) is uniformly bounded from above. Therefore, the RHS of (A.14) has the form of $-e_k^{*T} A_k e_k^* + K_{1_k} - 2e_k^{*T} B_k$, where $A_k \ge \beta I > 0$ and $K_{1_k} < M < \infty$. Furthermore, there exists a matrix $C \ge \alpha I > 0$ for some $\alpha > 0$ and a constant $0 < K < \infty$, for all $k \in Z_+$ such that

$$
-e_k^{*T} A_k e_k^* - e_k^{*T} B_k \le -e_k^{*T} C e_k^* + K. \qquad (A.16)
$$

If we denote the index k as $i-1$, and $k+N$ as $i+1$, then it is easy to see that (A.14) satisfies (A.10). Therefore, (A.12) can also be obtained. The rest of the proof is the same as that of Theorem 1. (q.e.d.)

3. *Proof of Theorem 3*

Introduce a primary Lyapunov function for the MGEKF as

$$
V_i(e_i) = e_i^T p_i^{*-1} e_i \qquad (A.17)
$$

where p_i^{*-1} is bounded from below by Assumptions 4 or 5 and p^* is a function of z_i^* and λ_i^* such that p_i^* is not correlated with v_i. Now a secondary Lyapunov function for the MGEKF is introduced as

$$
\bar{V}_i(\bar{e}_i) = \bar{e}_i^T m_i^{*-1} \bar{e}_i \qquad (A.18)
$$

where m_i^* is a function of z_{i-1}^* and λ_{i-1}^* such that m_i^* is independent of ω_{i-1}. Take the conditional expectation over $V_i(e_i) - \bar{V}_i(\bar{e}_i)$ for given $\bar{Y}_i = \{e_0, e_1, \ldots, e_{i-1}, \bar{e}_i\}$. Then from the hypothesis (3.35)

$$
E_{\bar{Y}_i}\{V_i(e_i) - \bar{V}_i(\bar{e}_i)\} = K_{1_i} - (1 - \rho_{1_i}) E_{\bar{Y}_i}\{\bar{V}_i(\bar{e}_i)\} \qquad (A.19)
$$

where K_{1_i} is defined as

$$
K_{1_i} = \mathrm{tr}\,[E_{\bar{Y}_i}\{(k_i^* + \Delta k_i)^T p_i^{*-1} (k_i^* + \Delta k_i) \gamma_i\}] \qquad (A.20)
$$

and K_{1_i} is bounded from above by the hypothesis of the theorem. Note that (A.19) is similar to (A.4) of Theorem 1. Similarly, if the hypothesis (3.36) is satisfied, the conditional expectation over $\bar{V}_i(\bar{e}_i) - V_{i-1}(e_{i-1})$ for given $Y_{i-1} = \{e_0, e_1, \ldots, e_{i-1}\}$ satisfies an equation similar to (A.8). The remainder of the proof is the same as that given in Theorem 1.

(q.e.d.)

ASSUMED DENSITY FILTER WITH APPLICATION TO HOMING MISSILE GUIDANCE

86-2262

S.N. Balakrishnan*
University of Missouri-Rolla
Rolla, Missouri
and
J.L.Speyer**
The University of Texas at Austin
Austin, Texas

ABSTRACT

A maximum likelihood estimation method is developed for a class of problems where the dynamics are linear and the measurement function is nonlinear. In this method, called the assumed density filter(ADF), the form of the conditional probability density function(CPDF) is selected to be a function of a finite number of quantities. These quantities which describe the approximate shape of the CPDF around the mode are propagated through each measurement interval. At the measurement the CPDF is updated using Bayes theorem and its mode, computed numerically, is defined to be the best estimate of the state. The posteriori CPDF is then approximated by a Taylor series expansion about its mode to preserve the assumed functional form. The numerical results for a target-intercept problem indicate that the ADF is superior to the extended Kalman filter. Howe·er.the ADF has a negative range bias. It is analytically proved,with some approximations, that the maximum likelihood range estimates are smaller than the mean range estimates.

1. INTRODUCTION

Tactical weapon systems require accurate tracking of maneuverable vehicles such as submarines and airplanes. During the last several years, there has been an active interest in the development of sophisticated filtering algorithms for tracking with bearings-only as the observations. Mathematically,this problem can be described in an inertial rectangular coordinate frame by a linear dynamical model and a nonlinear discrete observation model or in an inertial polar coordinate frame by a nonlinear dynamical model and a linear discrete observation model. Satisfactory results for this class of problems have been difficult to obtain using current mechanizable filters because of the nonlinearity and the passive nature of the observations. As a result, considerable research has been going on to improve the existing methods in order to obtain better estimates.

2 APPROXIMATIONS IN NONLINEAR FILTERING THEORY

The target tracking problem is stochastic in nature. Analyses of stochastic problems are possible through statistical interpretations. In order to obtain mathematical expressions for the statistics, assumed to represent the best estimates of the states associated with a problem, knowledge of the underlying probability density function (PDF) is essential.

If the system dynamics and/or the measurement function are nonlinear,a finite set of statistics sufficient to describe the conditional probability density function (CPDF) is not available (1,2,3). Even if the initial states and the process noise are assumed Gaussian,the nonlinear dynamical system results in a non-Gaussian CPDF. Second,the propagation equation for the conditional mean consists of expectations of nonlinear functions that are very difficult to evaluate. Third,since the CPDF is not Gaussian, the system of equations to describe the conditional moments and,hence,the filter as a whole becomes coupled and infinite-dimensional.

To circumvent these difficulties,approximations have been attempted to realize estimation methods consisting of a finite number of equations. The Edgeworth series expansion has been used by Sorenson and Stuberrud(4) in developing a finite-dimensional filter for a discrete scalar system. The Edgeworth series (5) is described by an asymptotic expansion about a kernel and it consists of Hermite polynomials (5) and their coefficients are given by quasi-moments which are related to the central moments. Sorenson and Stubberud have chosen a Gaussian kernel for their applications and approximated the CPDF by an Edgeworth series truncated after the fourth term. Thus,the first four quasi-moments define the CPDF. The problem of expectations of nonlinear functions in the propagation equations has been handled by series expansions. The state and measurement equations were approximated with second-order perturbations using a Taylor series. Due to the nonlinear term in the measurements,the update equations across the measurements become very involved requiring further approximations. Note that the filter developed in this paperdoes not have such approximations for the nonlinear measurements. Furthermore,Sorenson and Stubberud have reported that the approximations associated with the nonlinear measurment term are of "critical importance" to stabilise the behavior of the second and the fourth moments and thereby, to the performance of their filter.

A similar method of parametrizing the CPDF has been reported by Willsky (6). He has used the study of random processes on the circle effectively to formulate,using Fourier series,a variety of nonlinear estimation problems, arising in the field of communications. He has discussed a few finite-dimensional approximations for a scalar continuous time problem (6). The results of approximating the CPDF with the first three coefficients of Fourier series when applied to a phase-tracking problem were found to be very poor. The reason according to Willsky was that the truncated terms of the series might not have been negligible. This was shown by an example assuming perfect knowledge of the phase.

An alternate method to completely neglecting the higher-order moments or coefficients is discussed by Kushner (3). Instead of truncation of the higher order moments, he has devised a method to replace them with lower order moments. The method, called the 'moment sequences', involves picking an 'n' parameter moment approximation to the CPDF. When moments of order higher than 'n' are encountered, they can be computed in terms of the first 'n' moments resulting in a better approximation than assuming them to be zero. Kushner has discussed the conditions to be satisfied for picking the moment sequence for a scalar problem. It is not clear from his paper as to how the moment sequences could be picked for a general multidimensional problem.

Some other formulations of parametrizing the CPDF with moments have been reported (1,7). An approximation has been made by assuming the CPDF as Gaussian and neglecting the even moments of order higher than four. The system and measurement nonlinearities in this method are carried to second-order. The resulting filter is known as the Gaussian second-order filter (1,7). A slightly different version of this filter has been derived assuming that the CPDF is almost symmetric and

*Assistant Professor, Aerospace Engineering
Member AIAA
**Professor Aerospace Eng. and Engineering Mech.
Fellow AIAA

concentrated near its mean. Such a basis allows for ignoring the fourth central moment,resulting in what is termed as the truncated second-order filter (1,7). The basic difference between the two filters is in the propagation equation for the second moment. The modified second-order filter (1) is the version of these two filters without the measurement term, which contains random noise, included in the covariance equaton. This is done in order to prevent the covariance from taking negative values. The most popular method in the applications of nonlinear filtering theory is the extension of the Kalman filter methodology (1,8), which is optimal for linear systems, to nonlinear problems through linear perturbation theory. The resulting filter is called the extended Kalman filter (1). However, it is not known how the statistics of the CPDF relate to the extended Kalman filter (EKF). Simulation results for a scalar problem using the second-order filters discussed here and the EKF have been reported by Schwartz and Stear (7). Their results showed no particular merits of any second-order or any distinct superiority of the second-order filters over the EKF.

Another technique to approximate the CPDF involves the idea of cumulants (6,9). The advantage of the cumulants over the moments is that while the higher order moments may not tend to zero ,it is reasonable to assume that the higher order cumulants tend to zero. Nakazimo (9) has assumed a Gram-Charlier expansion (5) for the CPDF characterized by cumulants. He has also derived the dynamical equations for the cumulants for a nonlinear continuous time problem that are infinite-dimensional in nature and discussed finite-dimensional approximations by truncation. Willsky (6) has referred to the possible approximation of the CPDF using cumulants for the nonlinear estimation problems in communication theory. Both have not discussed any numerical results.

3. A NEW FILTER FORMULATION AND ITS RELATION TO PREVIOUS WORK

All the approximate filters, described in Section 2, claim to estimate the conditional mean. A major difficulty in the estimation of the conditional mean is the computation of the normalizing constant of the CPDF at the measurement update for the nonlinear problem (1). In this study,new filter structures which are more complex but mechanizable are proposed. The conditional mode is assumed to be the closest representation of the state, thus, eliminating the normalizing constant from the computations. Also,there is no approximation to the nonlinear measurement function as in the EKF. The basic idea is to choose the form of the CPDF to be a function of a finite number of quantities and to project these quantities through each measurement interval. These quantities describe the approximate shape of the CPDF around the mode. The vector which maximizes the approximate posterion CPDF at a measurement is defined to be the best estimate of the true state. This method is referred to in this paper as the assumed density filter (ADF). The ADF is applicable to a class of problems where the dynamics are linear and the measurement functions are nonlinear. The equations that define the ADF are developed and the ADF is applied to a homing missile problem. The results are explained and the inherent biases in certain formulations are indicated. The performance of the ADF is also compared to the widely-used EKF.

3.1 Assumed Form of The Density Function

In this section, an assumed form of the unnormalized CPDF is presented. Then, the procedure for processing measurements is developed, and the equations for propagating the quantities required for measurement processing are discussed. To avoid having to develop a theory for three-dimensional matrices, a mixed matrix-indicial notation is used. Hence, throughout this section, a repeated index denotes summation. The equations are developed in rectangular coordinates where the propagation equation for the approximate conditional mode is linear.

There are two considerations in the selection of the form of the PDF. First, the density function should be a function of a finite number of parameters. Second, the functional form of the density should be preservable during the processing of a measurement. These conditions can be accomplished by picking an exponential form for the density function and by writing the argument of the exponent in the form of a Taylor series about the mode and neglecting terms higher than a predetermined order Hence, the assumed form of the PDF of an n-state vector x is given by

$$\bar{p}(x,t) = C_1 \exp[-f(x,t)] \tag{1}$$

where C_1 is a normalizing constant and f(x,t) is a non-negative valued function of x. The function f(x,t) is assumed expandable in a Taylor series about the mode of the PDF, m, and expressed as

$$f(x,t)=f(m,t)+f_x(m,t)(x-m)+\frac{1}{2!}(x-m)^T f_{xx}(m,t)(x-m)$$

$$+\frac{1}{3!}(x-m)^T f_{xxx_j}(m,t)(x_j-m_j)(x-m) \tag{2}$$

$$+\frac{1}{4!}(x-m)^T f_{xxx_jx_k}(m,t)(x_j-m_j)(x_k-m_k)(x-m)+\ldots..$$

Here, for example, the term f_{xxx} denotes the partial derivative of the matrix f_{xx} with respect to the component state x_j Also, note that f and its derivatives are evaluated at the mode and are functions of the time, as is the mode.

The development of the ADF requires the equations for processing a measurement and the equations for propagating the terms f, f_x, f_{xx}, etc., to the next measurement time. However, at the mode, $p_x = 0$,so that $f_x(m,t)=0$ everywhere. In addition, it turns out that the term f(m,t) does not have any effect on the processing of the measurements; hence, it does not have to be propagated between measurements.

The remaining items needed for the development of the filter are the system dynamics and the measurement-state relation The system is assumed to be linear. Hence, the dynamics are governed by the equation

$$\dot{x}(t) = Fx(t) + b(t) + w(t) \tag{3}$$

where x is the n-state vector, F is n×n matrix of constants, b is a time-varying n-vector control, and w is a Gaussian zero-mean white-noise process with a constant power spectral density Q and t denotes the time. The relationship between the measurement and the state is represented by the vector equation

$$z_l = h(x_l) + v_l \tag{4}$$

Here,z_l is the p-vector measurement, h is the p-vector known nonlinear functions of the state x_l, v_l is a p-vector Gaussian zero-mean sequence of random variables with variance V, and the subscript l denotes the time at which the measurement is made.

3.2 Update Equations

The approximate CPDF prior to the measurement z_l is given by

$$\bar{p}(x_l/Z_{l-1})=C_{2l}\, exp\,(-\frac{1}{2!}(x-\bar{m})^T\bar{f}_{xx}(x-\bar{m})$$

$$-\frac{1}{3!}(x-\bar{m})^T f_{xxx_j}(x_j-\bar{m}_j)(x-\bar{m})$$

$$-\frac{1}{4!}(x-\bar{m})^T\bar{f}_{xxx_jx_k}(x_j-\bar{m}_j)(x_k-\bar{m}_k)(x-\bar{m})-..)_l \tag{5}$$

where $C_2 = C_1\, exp(\overline{-f})\,\bar{m}$ is the apriori mode the bar denotes quantities evaluated at $\bar{m}$, and Z_{l-1} denotes the measurement history upto l-1.

In processing the measurement z_l, the approximate conditional density function is updated using Bayes theorem which leads to

$$\bar{p}(x_l/Z_l) = C_{3l}\exp[-f(x,z)_l)] \tag{6}$$

where C_3 is the posteriori normalizing constant which is not a function of x but only of Z_l and where

$$f(x,z) = \frac{1}{2}(z-h)^T V^{-1}(z-h) + \frac{1}{2!}(x-m)^T \bar{f}_{xx}(x-\bar{m})$$

$$+ \frac{1}{3!}(x-\bar{m})^T \bar{f}_{xxx_j}(x_j-\bar{m}_j)(x-\bar{m}) \tag{7}$$

$$+ \frac{1}{4!}(x-\bar{m})^T \bar{f}_{xxx_jx_k}(x_j-\bar{m}_j)(x_k-\bar{m}_k)(x-\bar{m}) + \ldots$$

At this point, Eq.(6) is maximized with respect to x_l to obtain the posterior mode $\hat{m}_l$ which is the maximum likelihood estimate of the state. For some problems, the maximization can be carried out analytically; however,if this is not possible, a numerical method such as the Newton-Raphson(10)
method must be employed. Note that the mode does not depend on C_{3l}. Finally, the posterior conditional density function is expanded in a Taylor series about the posterior mode to obtain

$$\hat{p}(x; Z_l) = C_{4l}\,exp\,[-\frac{1}{2!}(x-\hat{m})^T \hat{f}_{xx}(x-\hat{m})$$

$$-\frac{1}{3!}(x-\hat{m})^T \hat{f}_{xxx_j}(x_j-\hat{m}_j)(x-\hat{m})$$

$$-\frac{1}{4!}(x-\hat{m})^T \hat{f}_{xxx_jx_k}(x_j-\hat{m}_j)(x_k-\hat{m}_k)(x-\hat{m}) + \ldots]_l \tag{8}$$

where $C_{4l} = C_{3l}\,exp\,(-\hat{f})$ and the carat denotes a quantity evaluated at $\hat{m}$. Note that the derivatives of f consist of the measurement z_l.

The functional form of the density function after the measurement, Eq.(8) , after the approximations is the same as that before the measurement, Eq.(5). Hence, the functional form has been preserved.

3.3 Propagation Equations

The differential equations for propagating the mode and the necessary derivatives of f are derived by repeated differentiation *of Eq.(1), Kolmogorov's equation (1),*

$$p_t = -p\ tr\ (F) - p_x^T(Fx+b) + \frac{1}{2}Q_{jk}p_{x_kx_j}, \tag{9}$$

and by use of the fact that $p_x = 0$ at the mode. The argument t of x ,p,p_x and p_{xx} have been dropped for convenience. For a Gaussian probability distribution, the derivative f_{xx} is the inverse of the covariance matrix. Hence, to be able to compare the results of the ADF with those of the EKF, the equations for propagating f_{xx} is replaced by an equation for propagating $f_{xx}^{-1} \equiv P$. Only, the equations through f_{xxx} are presented because this is the highest-order derivative present in the differential equation for the mode.

In view of the above discussion, the equation for propagating the mode is given by

$$\dot{m} = Fm + b - \frac{1}{2}P\ (Q_{jk}f_{x_kx_j x})^T \tag{10}$$

Then, the equations for propagating P can be written as

$$\dot{P} = \frac{1}{2}P\ f_{xxx_r}\ P_{rs}Q_{jk}f_{x_kx_jx_s} + FP + PF^T$$

$$+ \frac{1}{2}P\,[Q_{jk}(f_{x_jx}f^T_{x_kx} + f_{x_jx}f^T_{xx} + f_{x_kx}f^T_{x_jx} \tag{11}$$

$$-f_{x_jx_kxx})]P - \frac{1}{2}PQ_{jk}f_{x_kx_j}$$

Note that P must be inverted in order to obtain the f_{xx} terms needed to perform this integration. Finally, the differential equation for f_{xxx} is the following:

$$\dot{f}_{xxx_r} = -3f_{xxx_s}F_{sr} + \frac{1}{2}f_{xxx_r}Q_{jk}f_{x_kx_j} + \frac{1}{2}(f_{x_jx}f^T_{x_kx}$$

$$+f_{xx}f_{x_kx_j} + f_{x_kx}f^T_{x_jx} - f_{xxx_kx_j})P_{sm}Q_{jk}f_{x_kx_jx_m}, \tag{12}$$

$$r = 1,2,\ldots,n$$

where the fifth-order terms have been discarded.

The boundary conditions for the propagation equations are the values of $\hat{m}$,$\hat{P}$ and $\hat{f}_{xxx}$ obtained after processing the last measurement. These propagation equations are then integrated upto the _ next _ measurement time to obtain the values of $\bar{m}$, $\bar{P}$, and $\bar{f}_{xxx}$ needed for the next measurement update.

3.4 A Second-Order Assumed Density Filter

A second-order ADF is developed in order to to illustrate the adaptive nature of the ADF. Note that in the terminology of the ADF, the EKF is a second-order filter because of the linearized measurement-state relation and the assumption of a Gaussian probability distribution.

If the Taylor series expansion is truncated after the second-order terms, the exponent of the posterior density function, Eq.(7), becomes

$$f(x) = \frac{1}{2}(z-h)^T V^{-1}(z-h) + \frac{1}{2!}(x-\bar{m})^T \bar{P}^{-1}(x-\bar{m}) \tag{13}$$

where the substitution $\bar{f}_{xx} \equiv \bar{P}^{-1}$ has been made. Then, the posteriori mode occurs when $f_x = 0$ or when

$$[-(z-h)^T V^{-1}h_x + (x-\bar{m})^T \bar{P}^{-1}]_{x=\hat{m}} = 0 \tag{14}$$

Finallv, if Eq.(13) is expanded in a Taylor series about $\hat{m}$ and terms higher than second-order are dropped (see the discussion before Eq.(8)) , the following result is obtained.

$$\hat{P} = [h_x^T V^{-1}h_x - (z_j-h_j)V^{-1}_{jk}h_{kxx} + \bar{P}^{-1}]^{-1}_{x=\hat{m}} \tag{15}$$

Here, the subscripts on V^{-1} refer to the elements of V^{-1} . Also, $\hat{P}$, *which is the inverse of the curvature at the likelihood point,* must be positive definite.

With regard to the propagation equations, Eqs.(10) and (11) reduce to

$$\dot{m} = Fm + b \tag{16}$$

$$\dot{P} = FP + PF^T + Q \tag{17}$$

The initial conditions are obtained from Eqs.(14) and (15).

Eq.(15) points out the adaptive nature of the ADF. The second-order term h_{kxx} allows the ADF to adapt to the measurement residuals. This feature can be very useful when the measurement uncertainty is inaccurately modeled or when the states are inaccurately initialised.

4. HOMING MISSILE-INTERCEPT PROBLEM

A specific application of interest for the filtering technique discussed in Section 3 is in the estimation of the states of a homing missile relative to a target and the target acceleration. A six-degree-of-freedom computer program (11), which simulates the interception of a maneuvering target by a bank-to-turn, short-range, air-to-air homing missile has been used to test the ADF and the EKF. The guidance scheme to compute the commanded missile acceleration has been based on an 'optimal' linear guidance law(12).

The launch geometry used in this analysis is described in Figure 1. For this inertial system, the Z_I axis is directed towards the earth's center, the X_I axis is aligned parallel with the missile's initial launch direction, and the Y_I axis is chosen to make the inertial system right-handed. The engagement geometry used in this analysis is characterized by the initial conditions: range, 3000 feet; altitude, 10,000 feet; aspect angle (θ_a), 120 degrees; and off-boresight angle (θ_b), 0.0 degree. The measurement and the state noise models and their statistics are the same as in (13). The initial state covariance are the same as in (14). The number of Monte Carlo trials is ten. A second-order ADF is used in all the numerical experiments.

In the first attempt to validate the ADF using the selected engagement geometry, the ADF fails to converge near the end of the trajectory because P, defined as the approximate covariance matrix, becomes indefinite. In trying to determine why the ADF behaves this way, it is found that towards the end, the residuals become large. This phenomenon occurs because the true range vector and the estimated range vector have their components in different quadrants. Although the error in the range is not very large, the difference between the meaurement which is made on the true trajectory and the filter-computed measurement is of the order of 180 degrees. Consequently, the optimization process does not converge. The optimization process is terminated on the condition that the changes to the azimuth and elevation angles become smaller than 10^{-6} . At this point, the curvature of the likelihood function need not be positive- definite and as a result P defined as the inverse of the curvature is also non-positive definite. Note that if the residual dependent term is not included in the updating of P, it always remains positive-definite as in the case of the EKF. In fact, in the formulations of the Gaussian and truncated second-order filters (1), this random term is dropped in order to avoid the covariance from becoming non-positive definite. In this analysis, however, the random term is retained so that it could add more information to the the covariance. To circumvent the convergence problem, the measurements that cause large non-converging residuals are discarded. The underlying reasoning is that if a measurement causes P to become indefinite, then it increases the uncertainty of the estimates which P represents. Consequently, it does not help the filtering process and ,therefore, is ignored. The propagated states and the covariance are used to continue to the next measurement time.

The error histories of the ADF and the EKF for various conditions are given in Figurès 2 through 8. The error is defined as the difference between the magnitudes of the true and the estimated range vectors. The magnitude of the estimated range is obtained by averaging over the ten Monte Carlo runs.

The range errors for the nominal case are given in Figure 2. The ADF tracks better than the EKF during most of the flight. However, it exhibits a negative range bias. The range error history when the initial range has perturbations of 500 feet in the positive and negative directions are given in Figures 3 and 4 respectively. In both cases, the performance of the ADF is superior to that of the EKF.

Since the filter does not know the actual measurement noise variance, a measurement mismatch, defined as the ratio of the actual value to the assumed measurement covariance in the filters is hypothesized in generating the measurements. The range errors of the simulations with a measurement mismatch of 0.1 is presented in Figure 5. The performance trends of both filters are similar to the nominal case in Figure 2 except that the ADF performs better during the later part of the flight. For the higher mismatch of ten, the performance of the ADF, as can be observed from Figure 6, is much worse than that of the EKF.

To simulate actual situations where uncertainties in both the states and the noise statistics can occur, experiments are made with perturbations to the initial states and measurement mismatches. The range error history with an initial state error of 500 ft and a measurment mismatch of 0.1 is presented in Figure 7. With the same intial error, and a measurement mismatch of ten, the results in the range errors are given in Figure 8. The ADF tracks better than the EKF in both cases. From these numerical experiments, it can be seen that the ADF shows a high sensitivity to a higher measurement mismatch. This can, however, be alleviated by adaptively estimating the measurement noise covariance (15,16). The EKF performs well when the noise statistics are uncertain. However, its response to imperfect initial conditions is poor.

5. BIASES IN THE MAXIMUM LIKELIHOOD FILTERS

As discussed in Section 4, the ADF fails to converge near the end of the trajectory because P,defined as the approximate posteriori state error covariance,becomes indefinite. Investigation of the history of the range errors shows that the failure of convergence is always preceded by the filter computed range approaching zero when the actual range is far from zero. This phenomenon leads to the conjecture that there might be a negative bias in the range computed by the ADF.

This section deals with the analysis of the bias in the maximum likelihood estimates of the states of a system. Two inequalities relating the ranges from different methods are established. They are

$$1)\ \hat{R}_m \le \hat{R} \tag{18}$$

where $\hat{R}_m$ is the unconditional maximum likelihood range estimate of two Gaussian variables x and y and $\hat{R}$ is the unconditional range obtained from the same Gaussian distribution.

$$2)\ \hat{R}_{m_i} \le \hat{R}_i \le \hat{R}_{M_i} \tag{19}$$

where $\hat{R}_{m_i}$ is the conditional maximum likelihood range of two random variables x_i and y_i with a CPDF $p(x_i, y_i/Z_i)$, $\hat{R}_i$ is the conditional mean of the same CPDF and $\hat{R}_{M_i}$ is the conditional maximum likelihood range of $p(R_i, \theta_i/Z_i)$, the CPDF of {R,θ } obtained from the CPDF of {x,y}.

5.1 Relationship Between The Mean Range And The Maximum Likelihood Range For An Unconditional Density Function

The relative magnitudes of the mean range and the maximum likelihood range of the two Gaussian random variables x and y are compared in this section. Expressions for the mean range and the maximum likelihood range are obtained and Minkowski's inequality (18) is used to prove that the mean range is always equal to or greater than the maximum likelihood range.

Assume that the joint PDF of two random variables x and y is given by

$$p(x,y) = C\ exp\left[-\frac{1}{2}(x-\bar{x}, y-\bar{y})\, P^{-1} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}\right] \tag{20}$$

where

$\bar{x}, \bar{y} \equiv$ the means of x and y respectively
$\bar{P} \equiv$ and the state error covariance matrix of x and y
$C \equiv$ a normalising constant.

The maximization of the joint PDF with respect to x and y leads to the minimization of the negative part of the argument of the exponent, called the likelihood function, L.

$$L \equiv (x-\bar{x}, y-\bar{y})\, \bar{P}^{-1} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}. \tag{21}$$

For comparison and later use, the likelihood function L is defined in terms of the range, R, and the azimuth angle,θ , as

$$L \equiv (R\cos\theta-\bar{x}, R\sin\theta-\bar{y})\, \bar{P}^{-1} \begin{bmatrix} R\cos\theta-\bar{x} \\ R\sin\theta-\bar{y} \end{bmatrix}. \tag{22}$$

where $R = (x^2+y^2)^{1/2}$ and $\theta = \tan^{-1}(y/x)$.

The values for the range, $\hat{R}_m$,and the azimuth angle,θ_m ,that minimize L are obtained by setting the partial derivatives of L with respect to R and θ to zero,to yield

$$x_m = \hat{R}_m \cos\hat{\theta}_m = \bar{x} \text{ and}$$

$$y_m = \hat{R}_m \sin\hat{\theta}_m = \bar{y} \tag{23}$$

where x_m and y_m are the values of x and y which minimize L.Also,

$$\hat{R}_m = (x_m{}^2 + y_m{}^2)^{1/2}. \tag{24}$$

The bias in the maximum likelihood estimate of the range $\hat{R}_m$ is shown by comparison with that of the mean value of the range R which is given by

$$\hat{R} = E[R] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x^2+y^2)^{1/2} p(x,y)\, dx\, dy \tag{25}$$

where E(.) is the unconditional expected value operator. Minkowski's inequality (18) can be used to state the following inequality.For $x \neq y$,

$$(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x^2+y^2)^{1/2} p(x,y)\,dx\,dy)^2 \geq (\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, p(x,y)\,dx\,dy)^2$$

$$+(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, p(x,y)\,dx\,dy)^2 . \qquad (26)$$

The terms on the right hand side of Eq.(26) are the mean values of x and y. That is,

$$\int_{-\infty}^{\infty} x\, p(x,y)\,dx\,dy \equiv E[x] = \bar{x} \qquad (27)$$

and

$$\int_{-\infty}^{\infty} y\, p(x,y)\,dx\,dy \equiv E[y] = \bar{y}. \qquad (28)$$

With the aid of Eqs.(26),(27),and (28),the inequality statement can be rewritten as

$$(E[R])^2 \geq \bar{x}^2 + \bar{y}^2 = \hat{R}_m^2 . \qquad (29)$$

Since the range is always positive,Eq.(29) implies that

$$\hat{R} = E[R] \geq \hat{R}_m \geq 0 . \qquad (30)$$

5.2 *Approximations To The Posteriori Conditional Density Function*

With the assumption that the measurement noise variance is very small (a reasonable practical assumption) the conclusion of Eq. (30) is shown to be valid for the posteriori PDF conditioned on the measurement. In order to compute the conditional mean range, an expression for the posteriori CPDF is required. Assuming that the apriori CPDF of x_i andy_i and the measurement noise distribution at time i are known, the posteriori CPDF can be computed using Bayes rule. In doing so, however, the evaluation of $p(z_i/Z_{i-1})$, an integral which is a normalising constant is difficult. Consequently, Laplace's technique (17) is used to approximate $p(z_i/Z_{i-1})$ with the assumption that the measurement variance V is very small. To facilitate easier computation of the conditional mean range, the CPDF is obtained in polar coordinates R and θ.

Assume that the apriori joint PDF of the state variables x and y at stage i, conditioned on the measurement history Z_{i-1} upto stage (i-1),is Gaussian with means $\bar{x}$ and $\bar{y}$ and is given by

$$p(x_i,y_i/Z_{i-1}) = \frac{1}{2\pi\|\bar{P}_i\|}\, exp[-\frac{1}{2}(x_i-\bar{x}_i, y_i-\bar{y}_i)\, \bar{P}_i^{-1} \begin{bmatrix} x-\bar{x}_i \\ y-\bar{y}_i \end{bmatrix}] \qquad (31)$$

where

$$\bar{P}_i^{-1} \equiv S_i = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} . \qquad (32)$$

The PDF of v_i , the measurement noise in Eq.(4), is expressed as

$$p(v_i) = \frac{1}{(2\pi\,\|V\|)^{1/2}}\, exp(-\frac{1}{2} v_i^2 V^{-1}) . \qquad (33)$$

In order to make the computation of the mean range simpler , the CPDF of the rectangular coordinates x_i and y_i is first transformed to the polar coordinates R_i and θ_i . The result is

$$p(R_i,\theta_i/Z_{i-1}) = \frac{1}{2\pi\|\bar{P}_i\|}\, R_i\, exp[-\frac{1}{2}(\alpha_i^T S_i \alpha_i)] \qquad (34)$$

where

$$\alpha_i = \begin{bmatrix} R_i\cos\theta_i - \bar{x}_i \\ R_i\sin\theta_i - \bar{y}_i \end{bmatrix} ,$$

In processing the measurement z_i ,the CPDF is updated using Bayes' ruleas

$$p(R_i,\theta_i,Z_i) = \frac{p(z_i/R_i,\theta_i,Z_{i-1})\, p(R_i,\theta_i/Z_{i-1})}{p(z_i/Z_{i-1})} \qquad (35)$$

By examining on the right hand side term by term,it can be noted that $p(R_i,\theta_i/Z_{i-1})$ already known. From the joint PDF of z_i and v_i, the marginal CPDF $p(z_i/Z_{i-1})$ can be reduced as (19)

$$p(z_i/Z_{i-1}) = \int_{-\infty}^{\infty} p((z_i-v_i)/Z_{i-1})\, p(v_i)\, dv_i$$

$$= \frac{C_{1i}}{(2\pi\|V\|)^{1/2}} \int_{-\infty}^{\infty} [\frac{1}{a_i(\phi)} + \qquad (36)$$

$$\frac{b_i(\phi)}{a_i^{3/2}(\phi)} \exp(\frac{b_i^2(\phi)}{2a_i(\phi)})\, ((\frac{\pi}{2})^{1/2} + \Delta_i)]\, exp[-\frac{1}{2} v_i^2 V^{-1}]\, dv_i$$

where $\phi = z_i - v_i$ and $\Delta_i = \int_{-\frac{b_i}{a_i^{1/2}}}^{0} exp[\frac{x^2}{2}]\, dx$ ...

Also,

$$a_i(\theta) = s_{11}\cos^2\theta + 2s_{12}\sin\theta\cos\theta + s_{22}\sin^2\theta \text{ and}$$

$$b_i(\theta) = \cos\theta(s_{11}\bar{x} + s_{12}\bar{y}) + \sin\theta(s_{12}\bar{x} + s_{22}\bar{y}) \qquad (37)$$

The integral on the right hand side of Eq.(36) is approximated by using Laplace's method (17). The idea is to use the assumption that the measurement variance is small and, therefore, V^{-1} is large. The main contribution to the integral, then, comes from the region where the dominating term, $\exp(-\frac{1}{2} v_i^2 V^{-1})$ is maximum. The term multiplying the dominant term is expanded in a Taylor series about the maximum point and the integral is then evaluated term by term. To first order, the approximated value of $p(z_i/Z_{i-1})$ [Appendix A.1] is given by

$$p(z_i/Z_{i-1}) = C_{1i}\, K_{1i}(\theta_i) \qquad (38)$$

where

$$K_{1i} \equiv [\frac{1}{a_i(\theta)} + \frac{b_i^{3/2}(\theta)}{a_i^{3/2}(\theta)} exp[\frac{b_i^2}{2a_i}\frac{(\theta)}{(\theta)}]((\frac{\pi}{2})^{1/2} + \Delta_i)] . \qquad (39)$$

By the substitutions of Eqs.(34),(36), and (39) into Eq.(35), it can be shown that

$$p(R_i,\theta_i/Z_i) = \frac{R_i}{2\pi\|\bar{P}_i\|} \frac{\exp[-1/2(\alpha_i^T S_i \alpha_i + (z_i - \theta_i)^2 V^{-1})]}{(2\pi\|V\|)^{1/2} C_{1i} K_{1i}(\theta_i)} \qquad (40)$$

5.3 *Determination Of The Conditional Mean Range*

With the expression for the approximated CPDF in Eq.(40), $\hat{R}_i$, the conditional mean range at i, is computed in this section. Approximations have to be made to the integrals in the expression for $\hat{R}_i$ assuming that the measurement variance is very small. The mean value of the posteriori range is given by

$$\hat{R}_i = \int_0^{\infty} R_i\, \bar{p}(R_i/Z_i)\, R_i$$

$$= \int_{-\infty}^{\infty}\int_0^{\infty} R_i\, \frac{\bar{p}(z_i/R_i,\theta_i,Z_{i-1})\bar{p}(R_i,\theta_i/Z_{i-1})}{\bar{p}(z_i/Z_{i-1})}\, d\theta_i\, dR_i \qquad (41)$$

By the substitution of Eq.(40) in Eq.(41) it can be shown (after some manipulations) that R_i reduces to '(19)

$$\hat{R}_i = \frac{b_i}{a_i} + [\frac{1}{a_i}[exp(-b_i^2/2a_i) + (\frac{\pi}{2})^{1/2}] + \frac{b_i}{a_i} exp(b_i^2/2a_i)]/$$

$$(exp(-b_i^2/2a_i) + b_i/a_i^{1/2}((\frac{\pi}{2})^{1/2} + \Delta_i)) \qquad (42)$$

where the terms containing a_i and b_i are evaluated at $\theta_i = z_i$. For future comparison,it is observed that when $b_i^2/2a_i$ is large $\exp(-b_i^2/2a_i) \approx 0$ and Eq.(42) reduces to

$$\hat{R}_i = \left[\frac{b_i}{a_i} + \frac{1}{b_i(1+(\frac{2}{\pi})^{1/2}\Delta_i)}\right]_{\theta_i = z_i} \qquad (43)$$

By an asymptotic expansion for Δ_i (17) and neglecting terms containing $exp(-\frac{b_i^2}{2a_i})$, Δ_i can be approximated as

$$\Delta_i = \sqrt{2\pi}\,. \qquad (44)$$

By substitution of Δ_i from Eq.(44) into Eq.(43) , the expression for $\hat{R}_i$ is reduced to

$$\hat{R}_i \approx \frac{b_i}{a_i} + \frac{1}{3b_i}\,. \qquad (45)$$

Note that both terms on the right hand side of Eq.(45) are positive. The first term will be shown to be common to $\hat{R}_{m_i}$ and $\hat{R}_{M_i}$. However, comparison of the second term of $\hat{R}_i$ with that of $\hat{R}_{m_i}$ and $\hat{R}_{M_i}$ will establish the biases of $\hat{R}_{m_i}$ and $\hat{R}_{M_i}$.

5.4 *Determination Of The Range That Minimizes The Likelihood Function.*

An expression for the maximum likelihood range of $p(x_i, y_i/Z_{i-1})$ is obtained in this section for comparison with $\hat{R}_i$, given in Eq.(45). Maximization of the CPDF for $x_i$andy$_i$ amounts to minimization of the likelihood function which is defined as (19)

$$L = \frac{1}{2}[(z_i - \theta_i)^2 V^{-1} + R_i^2 a_i - 2R_i b_i + c_i]\,. \qquad (46)$$

where

$$c_i = s_{11}\bar{x}_i^2 + 2s_{12}\overline{xy}_i + s_{22}\bar{y}_i^2$$

The values of $\hat{R}_{m_i}$ and $\hat{\theta}_{m_i}$ which minimize L(in the process of maximizing the PDF) are obtained as (19)

$$\hat{R}_{m_i} = \left.\frac{b_i}{a_i}\right]_{\theta_i = \hat{\theta}_{m_i}} \qquad (47)$$

and

$$\hat{\theta}_{m_i} = z_i - V(R_i^2 a_i' - 2R_i b_i')\Big]_{\substack{R_i = \hat{R}_{m_i} \\ \theta_i = \hat{\theta}_{m_i}}}. \qquad (48)$$

where $a_i' \equiv \frac{\partial a_i}{\partial \theta_i}$ and $b_i' \equiv \frac{\partial a_i}{\partial \theta_i}$. Consistent with the argument in approximating $\hat{R}_i$, the measurement noise variance V is assumed to be very small. Consequently, the term containing V in Eq.(48) is neglected. The result is

$$\hat{\theta}_{m_i} \approx z_i\,. \qquad (49)$$

Eq.(49) leads to

$$\hat{R}_{m_i} = \left.\frac{b_i}{a_i}\right]_{\theta_i = z_i} \qquad (50)$$

By comparing the expression for $\hat{R}_{m_i}$ with that of $\hat{R}_i$ in Eq.(45), it can be concluded that $\hat{R}_{m_i}$,the conditional maximum likelihood estimate given by Eq.(50),is always smaller than $\hat{R}_i$,the conditionl mean range ,subject to the assumptions made.

5.5 A Method Of Modifying The Likelihood Estimate

Since the maximum likelihood estimate is biased,techniques for the reduction of the bias are developed. One method is determined by considering the apriori PDF transformed from x and y to R and θ which can be written as

$$p(R_i, \theta_i/Z_{i-1}) = \frac{1}{2\pi\|\bar{P}_i\|} R_i\, exp(-\frac{1}{2}(R_i^2 a_i - 2R_i b_i + c_i))\,. \qquad (51)$$

After processing the measurement, z_i ,the posteriori CPDF can be computed as

$$p(R_i, \theta_i/Z_i) = \frac{1}{p(z_i/Z_{i-1})} \frac{1}{2\pi\|\bar{P}_i\|} R_i\, exp(-\frac{1}{2}[(z_i - \theta_i)^2 V^{-1}$$

$$+ R_i^2 a_i - 2R_i b_i + c_i])\,. \qquad (52)$$

The transformed CPDF given by Eq.(52) is used to obtain the maximum likelihood estimates of R_i and θ_i instead of defining $\hat{R}_{m_i} = (x_{m_i}^2 + y_{m_i}^2)^{1/2}$ where x_{m_i} and y_{m_i} are the maximum likelihood estimates of the posteriori CPDF in x and y . It can be proved now that,$\hat{R}_{M_i}$,the range thus obtained is always greater than $\hat{R}_{m_i}$.

The maximization equations for $p(R_i\theta_i/Z_i)$ lead to (19)

$$\hat{R}_{M_i} = \left.\frac{b_i + (b_i^2 + 4a_i)^{1/2}}{2a_i}\right]_{\theta_i = \hat{\theta}_{M_i}} \qquad (53)$$

and

$$\hat{\theta}_{M_i} = z_i - V[R_i^2 a_i' - 2R_i b_i']\Big]_{\substack{R_i = \hat{R}_{M_i} \\ \theta_i = \hat{\theta}_{M_i}}}. \qquad (54)$$

The assumption that the measurement noise covariance, V, is small, is used in Eq.(54) to obtain an approximation to $\hat{\theta}_{M_i}$ as

$$\hat{\theta}_{M_i} \approx z_i\,. \qquad (55)$$

Substitution of Eq.(55) into Eq.(53) is made to yield

$$\hat{R}_{M_i} = \left.\frac{b_i + (b_i^2 + 4a_i)^{1/2}}{2a_i}\right]_{\theta_i = z_i}. \qquad (56)$$

The right hand side of Eq.(56) is now rewritten and expanded to show that $\hat{R}_{M_i}$ is always greater than $\hat{R}_{m_i}$.

$$\hat{R}_{M_i} = \frac{b_i}{2a_i} + \frac{b_i}{2a_i}(1 + 4\frac{a_i}{b_i^2})^{1/2}$$

$$= \frac{b_i}{2a_i} + \frac{b_i}{2a_i}(1 + \frac{1}{2}\frac{4a_i}{b_i} + O(\frac{a_i^2}{b_i^4})) \qquad (57)$$

where $O(\frac{a_i^2}{b_i^4})$ contains terms of the series of order equal to or higher than $(\frac{a_i^2}{b_i^4})$. This binomial expansion is valid for $\frac{4a_i}{b_i^2} < 1$ and this condition is usually satisfied in actual cases (Appendix A.2). Neglecting terms of order higher than $(\frac{a_i}{b_i^2})$ the range $\hat{R}_{M_i}$ can be approximated as

$$\hat{R}_{M_i} \approx (\frac{b_i}{a_i} + \frac{1}{b_i})\Big]_{\theta_i = \hat{\theta}_{M_i} = z_i}. \qquad (58)$$

This expression on the right hand side of Eq.(58) is always greater than $\hat{R}_{m_i}$ in Eq.(50) because a_i and b_i are always positive.

Comparison of $\hat{R}_{M_i}$ given by Eq.(58) with $\hat{R}_i$ in Eq.(45) shows that $\hat{R}_{M_i}$ is also biased. However, a positive bias as with $\hat{R}_{M_i}$ may be less detrimental to the performance of the filtering process than the negative bias of $\hat{R}_{m_i}$ because the range goes to zero in the homing missile problem.

6. CONCLUSIONS

In this study, the problem of obtaining estimates of the states of a linear system for the case where the measurement function is nonlinear has been considered. A new maximum likelihood filter has been formulated based on the Taylor series expansion of the posteriori conditional probability density function around its mode. The performance of the resulting assumed density filter is analysed by tracking a manuevering target with signals from a passive sensor for which only angle information is available. The numerical results show that the error histories of the new assumed density filter are better than that of the widely-used extended Kalman filter for initial conditions which are off-nominal. The assumed density filter is, however, negatively biased with respective to range.

The numerical results are corroborated by approximate analytical expressions for the conditional and unconditional estimates of the mean and maximum likelihood range.

APPENDIX A

This appendix deals with some derivations that are used in Section V.

1. Proofs To Show That $a(\theta)>0$ and $b(\theta)\geq 0$

Consider the likelihood function L in terms of x and y. It is given by

$$L = \frac{1}{2}\,(x-\bar{x}, y-\bar{y})\,S\,\begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix} \tag{A.1}$$

after dropping the subscripts for convenience. In Eq.(A.1), the matrix S, being the inverse of the apriori curvature at the maximum of the CPDF is positive. Given the quadratic form of the right hand side of Eq.(A.1), it can be concluded that L is always greater than or equal to zero. In terms of R and θ ,the likelihood function can be expressed as

$$L = \frac{1}{2}\,(R\cos\theta-\bar{x}, R\sin\theta-\bar{y})\,S\,\begin{bmatrix} R\cos\theta-\bar{x} \\ R\sin\theta-\bar{y} \end{bmatrix} \tag{A.2}$$

where R $=(x^2+y^2)^{1/2}$ and $\theta = \tan^{-1}(\frac{y}{x})$. The right hand side of Eq.(A.2) is expanded to yield

$$L = \frac{1}{2}\,(a(\theta)R^2 - 2b(\theta)R + c) \tag{A.3}$$

where

$$a(\theta) = s_{11}\cos^2\theta + 2s_{12}\cos\theta\sin\theta + s_{22}\sin^2\theta \tag{A.4}$$

$$b(\theta) = \cos\theta(s_{11}\bar{x}+s_{12}\bar{y}) + \sin\theta(s_{12}\bar{x}+s_{22}\bar{y}) \tag{A.5}$$

and

$$c \quad = s_{11}\bar{x}^2 + 2s_{12}\,\overline{xy} + s_{22}\bar{y}^2. \tag{A.6}$$

Note that $a(\theta)$ from Eq.(A.4) can be written in the following form:

$$a(\theta) = (\cos\theta, \sin\theta)\,S\,\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \tag{A.7}$$

Since $\cos\theta$ and $\sin\theta$ cannot both be zero at the same time, $a(\theta)>0$. By similar arguments, it can be shown that $c\geq 0$.

In order to prove that $b(0)\geq 0$ the likelihood function L is written as follows:

$$L = \frac{1}{2}\,a(0)(R-R_1)\,(R-R_2) \tag{A.8}$$

where

$$R_1 + R_2 = 2\,\frac{b(\theta)}{a(\theta)} \text{ and} \tag{A.9}$$

$$R_1R_2 = \frac{c}{a(\theta)} \geq 0 \tag{A.10}$$

Since both $a(\theta)$ and c are positive, R_1 and R_2 are such that they should satisfy one of the following conditions:

$$R_1 = 0 \quad R_2 \leq 0 \tag{A.11}$$

$$R_1 \leq 0 \quad R_2 = 0 \tag{A.12}$$

$$R_1 < 0 \quad R_2 < 0 \tag{A.13}$$

$$R_1 \geq 0 \quad R_2 \geq 0. \tag{A.14}$$

In addition to the restrictions on L, a, and c, note that in addition $R\geq 0$. This means that the value of R corresponding to the maximum value of L is also equal or greater than zero. Therefore, the only possible values are given by Eq.(A.14) . In order to satisfy Eq.(A.14), it can be observed from Eq.(A.9) that $b(\theta)\geq 0$.

2. Approximations Of $p(z_i/Z_{i-1})$

The aprioi PDF of z is given by

$$p(z_i/Z_{i-1}) = C_{1i}[T_1 + T_2] \tag{A.15}$$

where

$$T_1 = \frac{1}{(2\pi\|V\|)^{1/2}} \int_{-\infty}^{\infty} \frac{1}{a_i(\phi)}\, exp\,[-\frac{1}{2}u_i^2V^{-1}]\, du_i \tag{A.16}$$

$$T_2 = \frac{1}{(2\pi\|V\|)^{1/2}} \int_{-\infty}^{\infty} \frac{b_i(\phi)}{a_i^{3/2}(\phi)}\, exp\,(\frac{b_i^2(\phi)}{2a_i(\phi)})\exp(-\frac{1}{2}u_i^2V^{-1})du_i \tag{A.17}$$

and

$$C_{1i} = \frac{1}{2\pi\|P_i\|}\, exp\,(-\frac{1}{2}c_i) \tag{A.18}$$

and $\phi = z_i - u_i$. The integration of the term T_1 is considered first. The term $\exp(-\frac{1}{2}u_i^2V^{-1})$ is maximum at u_i =0 and becomes very small rapidly for any non-zero values of u_i because V is assumed to be very small. Note that for all values of θ, $a_i(\theta)$ is bounded. Under these conditions, an approximation to $p(z_i/Z_{i-1})$ is sought using Laplace's method (17). Laplace's method involves the expansion of the term multiplying the dominating exponential term in a power series and the evaluation of the resulting series of integrals term by term to any desired accuracy. In order to use the method the right hand side of Eq.(A.16) is expanded in a Taylor series about u_i=0. The result is

$$T_1 = \frac{1}{(2\pi\|V\|)^{1/2}} \int_{-\infty}^{\infty} [g(\phi)+g'(\phi)u_i+g''(\phi)u_i^2+....]$$

$$exp\,(-1/2u_i^2V^{-1})\, du_i \tag{A.19}$$

where $g(\phi) = \frac{1}{a_i(\phi)}$ and primes denote partial differentiation with respect to ϕ. All functions of ϕ are evaluated at $\phi = z_i$ Evaluation of the integrals in Eq (A.19) is carried out to yield

$$T_1 = g(\phi) + g''(\phi)V + O(V^2). \tag{A.20}$$

evaluated at $\phi = z_i$. Since V is small, the terms containing V are neglected and the approximated expression of T_1 is given by

$$T_1 \approx g(\phi)\Big]_{\phi=z_i} = \frac{1}{a_i(z_i)}. \tag{A.21}$$

In evaluating T_2, it is recognised that the term multiplying $exp(-1/2u_i^2V^{-1})$ also has an exponential term. However, $\exp(\frac{b_i^2(\phi)}{2a_i(\phi)})$ is bounded for all values of ϕ and it can be expanded in series. Each term in the expansion can be multiplied with $\frac{b_i(\phi)}{a_i^{3/2}(\phi)}$ and the resulting integrals can be evaluated as in the case of T_1. In other words, the term $\frac{b_i(\phi)}{a_i^{3/2}(\phi)} \exp(\frac{b_i^2(\phi)}{2a_i(\phi)})$ can be expanded in a Taylor series about $u_i = 0$ and the resulting expression for T_2 is

$$T_2 = \frac{1}{(2\pi\|V\|)^{1/2}} \int_{-\infty}^{\infty} [g_1(\phi) + g_1'(\phi)u_i + g_1''(\phi)u_i^2 + \ldots]$$

$$exp(-1/2u_i^2V^{-1})\,du_i \tag{A.22}$$

where

$$g_1 \equiv \frac{b_i(\phi)}{a_i^{3/2}(\phi)}\, exp(\frac{b_i^2(\phi)}{2a_i(\phi)}) \tag{A.23}$$

and the primes denote partial differentiation with respect to ϕ. All functions of ϕ are evaluated at $\phi = z_i$. Evaluation of the integrals in Eq.(A.22) yield

$$T_2 = g_1(\phi) + g_1''V + O(V^2). \tag{A.24}$$

As before with T_1, the second term is proportional to V, the third term is proportional to V^2 and so on. Neglecting the terms containing V, T_2 can be approximated as

$$T_2 \approx g_1(\phi)\Big]_{\phi=z_i} = \frac{b_i(z_i)}{a_i^{3/2}(z_i)}\, exp(\frac{b_i^2(z_i)}{2a_i(z_i)}). \tag{A.25}$$

By substitution of Eq.(A.21) and (A.25) into Eq.(A.15) gives an approximation for $p(z_i/Z_{i-1})$ is obtained as

$$p(z_i/Z_{i-1}) = C_{1i}\left[\frac{1}{a_i(z_i)} + \frac{b_i(z_i)}{a_i^{3/2}(z_i)}\, exp(\frac{b_i^2(z_i)}{2a_i(z_i)})\right] \tag{A.26}$$

3. *Arguments To Show* That $\frac{b_i^2(\theta)}{a_i(\theta)}$ Is "Usually" Greater Than Unity At The Maximum

The expression for $\frac{b_i^2}{a_i}$ at the maximum of the CPDF is given by

$$\frac{b^2}{a} = \frac{[cos\,\theta_m(s_{11}\bar{x} + s_{12}\bar{y}) + \sin\theta_m(s_{12}\bar{x} + s_{22}\bar{y})]^2}{s_{11}\cos^2\theta_m + 2s_{12}\cos\theta_m\sin\theta_m + s_{22}\sin^2\theta_m} \tag{A.27}$$

after suppressing the arguments and the subscripts of a,b,$\bar{x}$, and $\bar{y}$ for convenience. In order to analyse the magnitude of $\frac{b^2}{a}$, it is assumed that the values of the trigonometric functions of the apriori and the posterion estimates of the bearings are equal. That is,

$$\cos\theta_m \approx \cos\bar{\theta} \quad \text{and} \quad \sin\theta_m \approx \sin\bar{\theta} \tag{A.28}$$

where

$$\bar{\theta} = \tan^{-1}(\bar{y}/\bar{x}).$$

By substitution of Eq.(A.28) into Eq.(A.27), an approximation for $\frac{b^2}{a}$ is obtained as

$$\frac{b^2}{a} = \frac{[cos\,\bar{\theta}\,(s_{11}\bar{x} + s_{12}\bar{y}) + \sin\bar{\theta}\,(s_{12}\bar{x} + s_{22}\bar{y})]^2}{s_{11}\cos^2\bar{\theta} + 2s_{12}\sin\bar{\theta}\cos\bar{\theta} + s_{22}\sin^2\bar{\theta}}$$

$$= s_{11}\bar{x}^2 + 2s_{12}\overline{xy} + s_{22}\bar{y}^2 \tag{A.29}$$

Without losing the generality, the magnitude of $\frac{b^2}{a}$ can be analysed by examining $s_{11}\bar{x}^2$. At the beginning, the initial range considered in a typical tracking or a homing missile problem is 3000 feet or more and , therefore, $\bar{x}$ is of the order of 10^3. Though the initial value of the covariance is assumed to be high, thereby, making s_{11} small, the product $s_{11}\bar{x}^2$ is usually greater than unity. As the problem proceeds, $\bar{x}$ becomes smaller in the homing missile problem. However, with more information through the measurements, the covariance matrix is also becomes smaller and, therefore, s_{11} bigger. Consequently, the magnitude of $\frac{b^2}{a}$ is greater than unity.

REFERENCES

1. Jazwinski, A., *"Stochastic Processes and Filtering Theory,"* Academic Press, 1970.

2. Kushner, H.J., "Nonlinear Filtering: The Exact Dynamical Equations Satisfied by the Conditional Mode," IEEE Transactions on Automatic Control, Vol. AC-12,No.3, June 1967.

3. Kushner, H.J., "Approximations to Optimal Nonlinear Filters," IEEE Transactions on Automatic Control, Vol. AC-12,No.5, October 1967.

4. Sorenson, H. W., and Stubberud, H. W., "Nonlinear Filtering by Approximation of the Posteriori Density," International Journal of Control, Vol. 8, NO. 1, pp. 33-51, 1968.

5. Kendall, M.G. and Stuart, A., " *The Advanced Theory of Statistics, Vol.I,"* Hafner Publishing Company, London 1969.

6. Willsky, A.S., "Fourier Series and Estimation on the Circle with Applications to Synchronous Communication-Part II: Implementation," IEEE Transactions on Information Theory, Vol. IT-20, No.5,September 1974.

7. Schwartz, L., and Stear, E., "A Computation Comparison of Several Nonlinear Filters," IEEE Transactions on Automatic Control, pp. 83-86, February 1968.

8. Kalman, R.E.,"A New Approach to Linear Filtering and Prediction Problems," Transactions of ASME, Journal of Basic Engineering, Vol.82 ,1960.

9. Nakamizo, T." On the State Estimation for Nonlinear Dynamical Systems," International Journal of Control, Vol.11, No.4, 1970.

10. Philips, G.M. and Taylor, P.J., *"Theory and Application of Numerical analysis,"* Academic Press, Newyork 1973.

11. Smith, R.C.,"A Six-Degrees of Freedom Computer Simulation of an Air-to-Air Missile Intercept," M.S. Thesis, The University of Texas at Austin, Texas, May 1979.

12. Fiske, P.H.,"Advanced Digital Guidance and Control Concepts for Air-to-Air Tactical Missiles," Report AFATL-TR-77-130, Air Force Armament Laboratory, Armament Development and Test Center, Eglin Air Force Base, Florida 32542.

13. Speyer, J.L. and Hull, D.G., "Comparison of Several Extended Kalman Filter Formulations for Homing Missile Guidance," Proceedings of the AIAA Guidance and Control Conference, Danvers, Massachusetts, August 1980.

14. Balakrishnan, S.N. and Speyer, J.L., "A Coordinate-Transformation Based Filter For Improved Target Tracking," Proceedings of The 23rd IEEE Conference on Decision and Control, Las Vegas, Nevada, 1984.

15. Myers, K.A.,"Filtering Theory Methods and Applications to Orbit Determination Problem for Near Earth Satellites," AMRL Report 1058, The University of Texas at Austin, Texas, 1973.

16. Greenwell, W.,"Adaptive Noise Estimation and Guidance for Homing Missiles," ,M.S.Thesis, The University of Texas at Austin, Texas 1982.

17. Carriere, G.F., Krook, M., and Pearson, C.E., *"Functions of a Complex Variable :Theory and Technique,"* McGraw-Hill Book Company 1966.

18. Hardy, G.H., Littlewood, J.E., and Polya, G., " *Inequalities* ," Cambridge University Press, 1934.

19. Balakrishnan, S.N., "Development Of Two Maximum likelihood Filters An Their Applications To Tracking Problems," Ph.D. Dissertation, Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, Texas, May 1984.

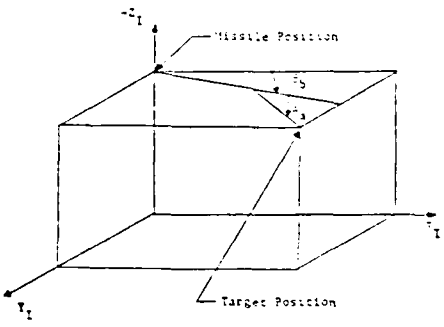


Figure 1. Launch Geometry

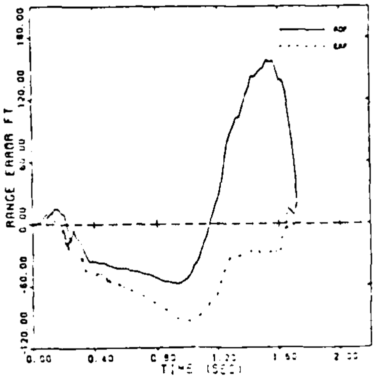


Figure 2. Range Error History
(No Perturbations)

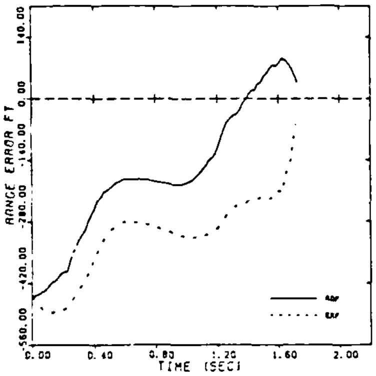


Figure 3. Range Error History
(I = 500 ft.)

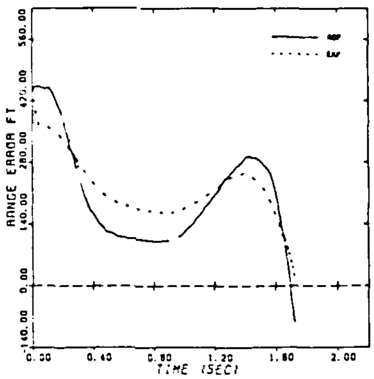


Figure 4. Range Error History
(I = -500 ft.)

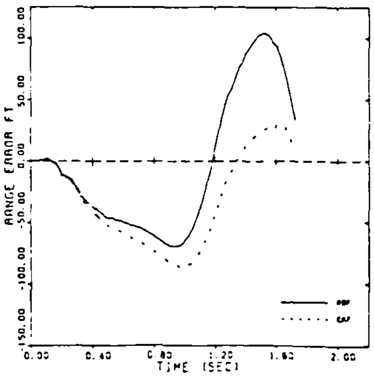


Figure 5. Range Error History
(Mismatch = 0.1)

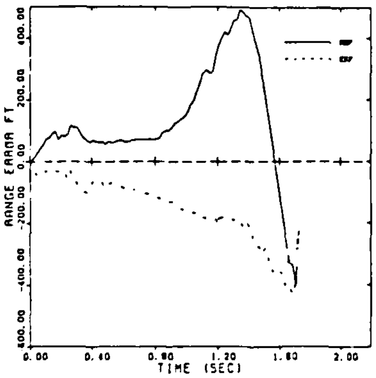


Figure 6. Range Error History
(Mismatch = 10)

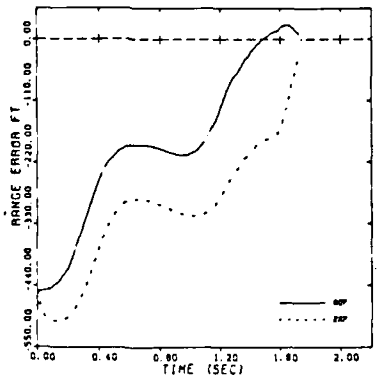


Figure 7. Range Error History
(Mismatch = 0.1, I = 500 ft.)

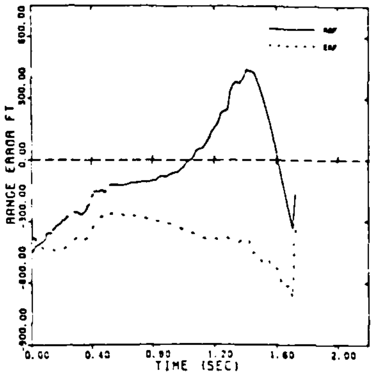


Figure 8. Range Error History
(Mismatch = 10, I =500 ft.)

Coordinate-Transformation-Based Filter for Improved Target Tracking

S. N. Balakrishnan*
University of Missouri—Rolla, Missouri
and
Jason L. Speyer†
The University of Texas at Austin, Texas

A maximum likelihood estimation method is developed for applications to the target tracking problem based on bearings-only observations from a single observer. The method involves propagation of states in rectangular coordinates in which the linear dynamics permit a closed form solution. At the measurement times the states are converted to a special polar coordinate system in which the measurement is modeled as linear in the transformed state and updated using the Kalman methodology. The coordinate transformation is chosen so that the direct transformation of the maximum likelihood estimate is approximately preserved. The numerical experiments for a target-intercept problem are presented, which show that the performance of this coordinate transformation based filter is superior to that of the Cartesian system based extended Kalman filter. Approximate analytical results also corroborate the numerical results.

Introduction

TACTICAL weapon systems require accurate tracking of maneuverable vehicles such as submarines and airplanes. During the last several years there has been an active interest in the development of sophisticated filtering algorithms for tracking with bearings-only observations. Satisfactory results have been difficult to obtain using current mechanizable filters because of the nonlinearity and passive nature of the observations. In practical cases, the high level of uncertainty in the initial states of the submarine and the rapidly accelerating target in the missile-intercept problem make accurate estimation of the states even more difficult to accomplish. Considerable research has been going on to improve existing methods. The single dominant method used, in the applications of the nonlinear filtering methods to the tracking problems, has been the extended Kalman filter (EKF).[1-5]

An approach to achieving a better nonlinear estimator is to determine a state-space, which may be different from Cartesian coordinates, for which improved estimation occurs.[6-9,13] The aim of this paper is to develop a better and suitable maximum likelihood filter based on transformations of state spaces for application to a target-intercept problem. The formulated polar coordinate filter (PCF) uses a nonlinear transformation of the state spaces. However, the choice of the coordinate systems is such that the approximate conditional mode (which represents the estimates of the states) is, except for one state, unaltered by the nonlinear transformations. This is approximately true in a three-dimensional coordinate frame and exactly true in a two-dimensional frame. The development of the PCF, numerical results from the application of the PCF to a homing missile problem, and analyses are used to explain the results of the numerical experiments. In all analyses, the Cartesian based EKF has been included for comparison.

Development of the Polar Coordinate Filter

The basis for the PCF is the availability of two convenient mathematical descriptions of the tracking problem. If Cartesian coordinates are used, the state equation is linear and the corresponding measurement equation is nonlinear. If polar coordinates are selected, the model has a nonlinear state equation but a linear measurement equation. The idea behind the PCF is to exploit available linearity in both coordinate systems. This idea has been used in the past by Mehra[9] and Sammons[3] to solve tracking problems. The algorithms of Mehra[9] and Sammons[3] use the standard Kalman update and propagation formulas where a nonlinear transformation is used for the covariances. In contrast, the approximations used in the PCF are made directly to the conditional probability density function (CPDF), and the approximate conditional mode is assumed to represent the best estimate of the state. In all of the available work involving transformations of the state spaces,[6,7,8,9] it is tacitly assumed that the transformation between the approximate CPDFs preserves the approximate conditional mean. However, this is not the case. In the PCF, the coordinate system is chosen so that the approximate conditional mode is least affected by the nonlinear transformation.

System Model

The nine-element state vector describing the missile-target engagement contains a three-dimensional relative position vector, a three-dimensional relative velocity vector, and a three-dimensional target acceleration vector modeled as a first order Markov process.[1] The evolution of the state vector in the inertial frame is written in matrix notation as

$$x = Fx + b + w \tag{1}$$

where x is the state vector in an arbitrary inertial frame consisting of the position vector (x_R, y_R, z_R), and the velocity and acceleration components are represented by the six-element vector v. F is a 9×9 matrix of constants. The nine-element vector b contains the components of the missile acceleration vector in the inertial axes and is given as $b = [0, 0, 0, a_{x}, a_{y}, a_{z}, 0, 0, 0]^T$. The only non-zero components in the nine-element vector w correspond to the target acceleration components. w is a Gaussian zero-mean white noise process with a power spectral density Q.

The discrete nonlinear noise corrupted angle observations are expressed in rectangular coordinates as

$$z_1 = \tan^{-1}(y_R/x_R) + v_1$$

$$z_2 = \tan^{-1}[-z_R/(x_R^2 + y_R^2)^{1/2}] + v_2 \tag{2}$$

Received Aug. 12, 1985; revision received March 25, 1986. Copyright © American Institute of Aeronautics and Astronautics, Inc., 1986. All rights reserved.

*Assistant Professor, Department of Mechanical and Aerospace Engineering. Member AIAA.

†Professor, Department of Aerospace Engineering and Engineering Mechanics. Fellow, AIAA.

where z_{1i} and z_{2i} are the measurements, v_{1i} and v_{2i} are Gaussian zero-mean sequences of random variables with a variance V, and subscript i denotes the time at which the measurement is made.

Transformations of the Approximate Conditional Probability Density Functions

Since the system dynamics are linear, its approximate conditional mean and mode are propagated from stage $(i-1)$ to stage i by

$$\bar{x}_i = \Phi(t_i, t_{i-1})\hat{x}_{i-1} + \int_{t_{i-1}}^{t_i} \Phi(t_i, \tau)b(\tau)\mathrm{d}\tau \tag{3}$$

where $\Phi(t_i, t_{i-1})$ is the state transition matrix, $(\bar{\ })$ is the a priori conditional mode at t_i given Z_{i-1}, the measurement history up to $(i-1)$, and $(\hat{\ })$ is the a posteriori conditional mode at i given Z_i. The propagation equation for the approximate a priori state error covariances for x is given by

$$\bar{P}_i = \Phi(t_i, t_{i-1})P_{i-1}\Phi^T(t_i, t_{i-1}) + \int_{t_{i-1}}^{t_i} \Phi(t_i, \tau)Q\Phi^T(t_i, \tau)\mathrm{d}\tau \tag{4}$$

where P_{i-1} is the known approximate a posteriori conditional covariance at $(i-1)$. $\bar{P}_i$ indicates the inverse of the curvature of the CPDF around the mode. Therefore, before a measurement update, the approximate a priori CPDF of x_i is assumed to be of the form

$$\bar{p}(x_i/Z_{i-1}) = \bar{c}_1 \exp[-(1/2)\alpha_i^T\bar{P}_i^{-1}\alpha_i] \tag{5}$$

where $\alpha_i = [x_R - \bar{x}_R, y_R - \bar{y}_R, z_R - \bar{z}_R, s_i^T - \bar{s}_i^T]^T$, and $\bar{x}_R$, $\bar{y}_R$, and $\bar{s}$ are defined to be the a priori modes of x_R, y_R, z_R, and s, respectively. $\bar{c}_1$ is a normalizing constant.

The transformation from the Cartesian coordinate system to a polar coordinate system is given in a functional form by $y = g(x)$ where y includes the two measurement functions θ and ϕ. The choice of the other states is dictated by the need to *maintain* (at least approximately) the direct transformation of the maximum likelihood estimates between the CPDF's of x and y. In order to achieve this, one of the states in y is the cube of the range $R3$ and the other is s. The transformed state y is given by $y \triangleq [R3, \theta, \phi, s^T]^T$ and

$$\begin{aligned} g(x) = {} & [(x_R^2 + y_R^2 + z_R^2)^{3/2}, \tan^{-1}(y_R/x_R), \\ & \tan^{-1}(-z_R/(x_R^2 + y_R^2)^{1/2}), s^T]^T \end{aligned} \tag{6}$$

The choice of $R3$ will be motivated in discussing the transformation of the CPDF. The inverse transformation from y to x is given by $x = h(y)$ where $x \triangleq (x_R, y_R, z_R, s^T)^T$ and

$$\begin{aligned} h(y) = {} & (R3^{1/3}\cos\phi\cos\theta, R3^{1/3}\cos\phi\sin\theta, \\ & -R3^{1/3}\sin\phi, s^T)^T \end{aligned} \tag{7}$$

The CPDF of y_i can be obtained from that of x_i as[10]

$$\bar{p}(y_i/Z_{i-1}) = \bar{p}(x_i/Z_{i-1})|J^{-1}| \tag{8}$$

where $|J^{-1}|$, the determinant of the Jacobian of the transformation, is found to be $1/3 \cos\phi_i$. With this value of $|J^{-1}|$, the CPDF of y_i can be written in terms of $R3_i$, θ_i, ϕ_i, and s_i as

$$\begin{aligned} \bar{p}(y_i/Z_{i-1}) & = \tfrac{1}{3}\cos\phi_i\bar{c}_1\exp[-\tfrac{1}{2}(\beta^T\bar{P}_i^{-1}\beta_i)] \\ & \triangleq \tfrac{1}{3}\cos\phi_i\exp[-f(y_i)] \end{aligned} \tag{9}$$

where

$$\begin{aligned} \beta_i = {} & [R3_i^{1/3}\cos\phi_i\cos\theta_i - \bar{x}_R, R3_i^{1/3}\cos\phi_i\sin\theta_i - \bar{y}_R, \\ & R3_i^{1/3}\sin\phi_i - \bar{z}_R, (s_i - \bar{s}_i)^T]^T \end{aligned}$$

If the determinant of the Jacobian of the transformation between the x and y systems is constant, the modes of the approximate CPDFs of both systems can be related by the transformation $\bar{y}_i = g(\bar{x}_i)$. Since the determinant of the Jacobian is not constant, approximations to the conditional modes of the y system need to be determined. Since $\cos\phi_i$ multiplies the exponentials in Eq. (9), the approximation to the conditional mode for ϕ is of concern. The approximate conditional mode $\bar{\phi}_i$ is obtained by setting the partial derivative of $\bar{p}(y_i/Z_{i-1})$, with respect to ϕ_i, to zero as

$$[\partial\bar{p}(y_i/Z_{i-1})]/[\partial\phi_i] = (-\tan\phi_i - \beta_i^T\bar{P}_i^{-1}r_i)\bar{p}(y_i/Z_{i-1}) = 0 \tag{10}$$

where $r_i \triangleq \partial\beta_i/\partial\phi_i$. The usual Kalman update formulas for estimation are obtained if $\tan\phi_i$ were not present or neglected. In the homing missile problem, which motivates our work, the initial launch geometry is assumed to be coplanar in the x-y plane, which means that the initial value of ϕ is zero. Throughout a typical engagement, the magnitude of ϕ_i is not greater than 45 deg and the magnitude of $\tan\phi_i$ is less than unity. In comparison with the other terms in Eq. (10), which contain terms of the order of the range $(R3^{1/3} = R)$ or higher, $\tan\phi_i$ is negligible. With this approximation, the approximate conditional modes of $\bar{p}(y_i/Z_{i-1})$ are obtained as $\bar{y}_i = g(\bar{x}_i)$.

This particular choice of transformation variables from Eq. (6) approximately preserves the *conditional modes* of the initial and the transformed approximate CPDFs. There are some examples in the literature[3,5,7,9] where transformations of the variables of interest have been used to form new state spaces where the EKF has performed better. In all of these sets of state spaces the assumption of a Gaussian form for the CPDF does not allow the nonlinear transformations to preserve the approximate conditional means (which are used to represent the estimate of the states) in the transformed coordinate systems.

If the CPDF can be approximated by a Gaussian form for $R3_i$, θ_i, ϕ_i, and s_i, then the states can be updated using the Kalman rule[10] at the measurement. In order to accomplish this, the argument of the exponent in Eq. (9) is expanded in a Taylor series about the approximate conditional mode up to second order. In neglecting the higher-order terms, it is assumed that the CPDF is approximately Gaussian, with the mean approximately equal to the mode. The resulting approximated CPDF

$$\bar{p}(y_i/Z_{i-1}) = \bar{c}_2\exp[-\tfrac{1}{2}(\delta_i^T\bar{P}_{p_i}^{-1}\delta_i)] \tag{11}$$

where $\delta_i \triangleq y_i - \bar{y}_i$, $\bar{c}_2 \triangleq \tfrac{1}{3}\cos\bar{\phi}_i\bar{c}_1\exp[-f(\bar{y}_i)]$, and the *mn*th element of $\bar{P}_{p_i}^{-1}$ is given by

$$(\bar{P}_{p_i}^{-1})_{mn} = [\partial^2(f(y_i) - \log\cos\phi_i)]/[\partial(y_i)_m\partial(y_i)_n]|_{y_i = \bar{y}_i} \tag{12}$$

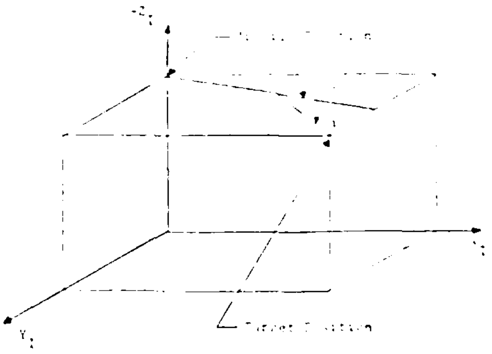


Fig. 1 Launch geometry.

In Eq. (11), $\overline{P}_{p}^{-1}$ represents the curvature of this CPDF at the a priori mode $\bar{y}_i$. The presence of the term exp(log cos ϕ_i) in Eq. (9) prevents a simple expression for $\overline{P}_{p_i}^{-1}$. Therefore, its effect is examined. The only term that log cos ϕ_i affects is $(\overline{P}_{p}^{-1})_{33}$, and it is given by

$$(\overline{P}_{p_i}^{-1})_{33} = \frac{\partial^2}{\partial\phi^2}(-\log(\cos\phi_i) + f(y_i))\Big|_{y_i=\bar{y}_i}$$

$$= \sec^2\overline{\phi}_i + R\overline{3}_i^{-2}F \qquad (13)$$

where F is a function of $\overline{P}_{p_i}^{-1}$, $\sin\overline{\phi}_i$, and $\cos\overline{\phi}_i$. Note that the term $\sec^2\overline{\phi}_i$ in Eq. (13) varies between 1 and 2 as $\overline{\phi}_i$ changes from zero to 45 deg. Compared to the other terms, which have a magnitude of the order of range, $\sec^2\overline{\phi}_i$ is small and therefore neglected. Since the expression for $\overline{P}_{p_i}^{-1}$ in Eq. (11) can be reduced to

$$\overline{P}_{p_i}^{-1} = (g_{x_i}\overline{P}_{x_i}g_{x_i}^T)^{-1} \qquad (14)$$

where $g_{x_i} \triangleq \partial g(x_i)/\partial x_i$ is evaluated at $x_i = \bar{x}_i$, the approximate CPDF represented by Eq. (11) is Gaussian.

In the coordinate systems employed by Mehra[9] and Sammons,[3] $y' \triangleq [R1, \theta, \phi, s]$ where $R1 \triangleq (x_R^2 + y_R^2 + z_R^2)^{1/2}$. The transformed CPDFs have the functional form

$$\bar{p}(y_i'/Z_{i-1}) = \varepsilon(y_i')\exp(f(y_i')) \qquad (15)$$

since the determinant of the Jacobian for their transformations is a function of y_i'. Furthermore, the determinate of the Jacobian is not even approximately justifiable as a constant since it contains the range. When the argument of the exponential is approximated up to second order, the resulting approximated CPDF is of the form

$$\bar{p}(y_i'/Z_{i-1}) = \varepsilon(y_i')\exp\{-1/2(y_i' - \bar{y}_i)^T\overline{P}_{p_i}'^{-1}(y_i' - \bar{y}_i')] \qquad (16)$$

which is clearly not Gaussian. Consequently the applications of Kalman methodology in their algorithms are not valid.

Updating the Conditional Density Functions and Estimation in Polar Coordinates

The a priori CPDF for $R3_i$, θ_i, ϕ_i and s_i given by Eq. (11) can be updated by using Bayes' rule.[10] The a posteriori CPDF, after some manipulation, can be reduced to a Gaussian CPDF

$$\bar{p}(y_i/Z_i) = c_3\exp[-1/2(y_i - \hat{y}_i)\hat{P}_{p_i}^{-1}(y_i - \hat{y}_i)] \qquad (17)$$

The objective is to update the conditional mode as if it were the conditional mean with the curvature at the conditional mode used as the inverse of the conditional covariance. The update equations used to process a measurement are given by the Kalman rules as

$$\hat{y}_i = \bar{y}_i + P_{p_i}H_p^TV_i^{-1}(z_{1_i} - \overline{\theta}_1, z_{2_i} - \overline{\phi}_i)^T \qquad (18)$$

$$\hat{P}_{p_i} = (\overline{P}_{p_i}^{-1} + H_{p_i}^TV_i^{-1}H_{p_i})^{-1} \qquad (19)$$

H_p is a constant vector or partial derivative of the measurement with respect to y_i evaluated at $\bar{y}_i$, and c_3 is a normalizing constant. There is no further approximation involved in the update process.

After the update in the polar coordinate system, the conversion back to the Cartesian system is obtained by tracing the same steps and making similar approximations to the value of ϕ_i, as before. The PCF is computed by the propagation Eqs. (3) and (4), transformation Eqs. (6) and (14), filter update Eqs. (18) and (19), and transformation back to the x system given by $\hat{x}_i = h(\hat{y}_i)$. The state covariance of the x system is obtained in a manner similar to Eq. (14), where

$$\hat{P}_{x_i} = g_{x_i}^T\hat{P}_{p_i}g_{x_i}^{-1} \qquad (20)$$

Numerical Results

A six-degree-of-freedom computer program,[11] which simulates the intercept of a maneuvering target by a bank-to-turn, short-range, air-to-air homing missile has been used to test the PCF and the EKF. The guidance scheme that computes the commanded missile acceleration is based on an "optimal" linear guidance law.[1]

The launch geometry used in this analysis is described in Fig. 1. For this inertial system, the Z_I axis is directed toward the Earth's center, the X_I axis is aligned parallel to the missile's initial launch direction, and the Y_I axis is chosen to make the inertial system right handed. The engagement is characterized by the initial conditions: range, 3000 ft; altitude, 10,000 ft; aspect angle (θ_a), 120 deg; and off-boresight angle (θ_b), 0.0 deg. The number of Monte Carlo trials is ten. The range-dependent measurement noise model and the state noise model and their statistics are the same as Ref. 12. The diagonal elements of the initial state covariances are 10^4 ft^2, corresponding to x_R, 10^7 ft^2, to y_R and to z_R, 100 ft^2/sec^2 for the velocity components, and 10 ft^2/sec^4 to the target acceleration components. The off-diagonal elements are zero.

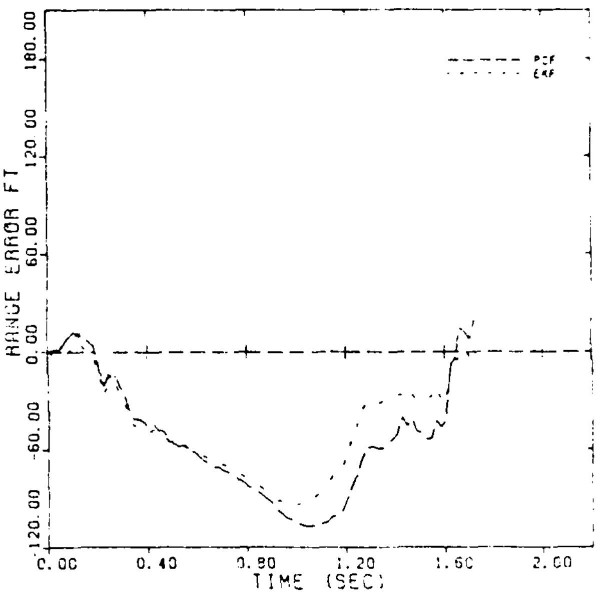


Fig. 2 Range error history (no perturbations).

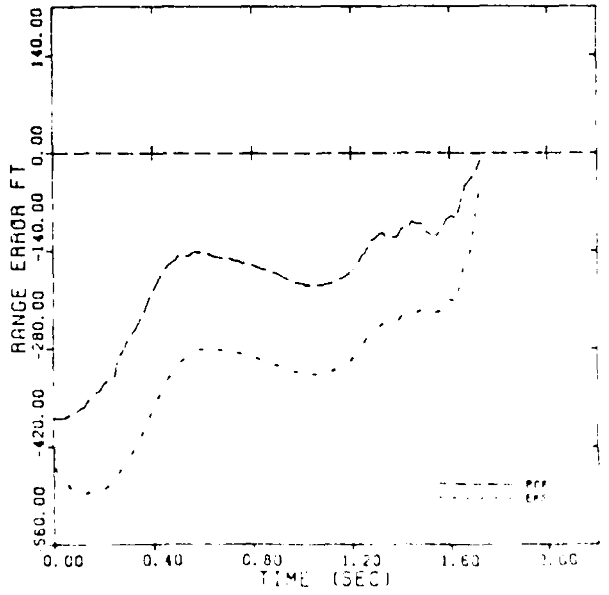


Fig. 3 Range error history (l = 500 ft).

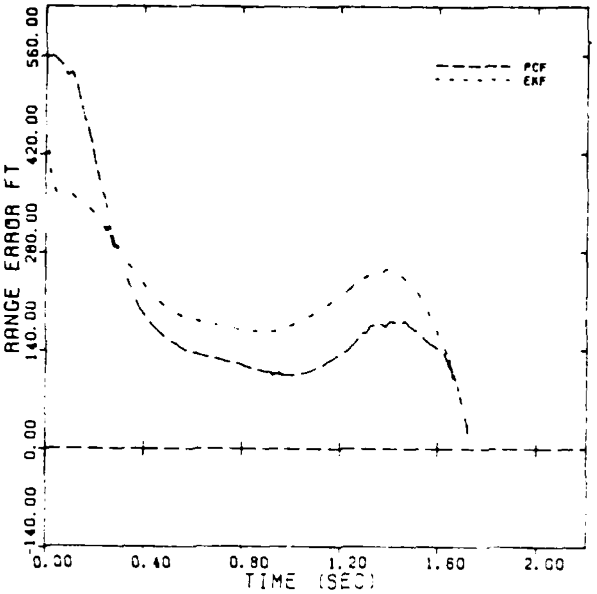


Fig. 4 Range error history ($I = -500$ ft).

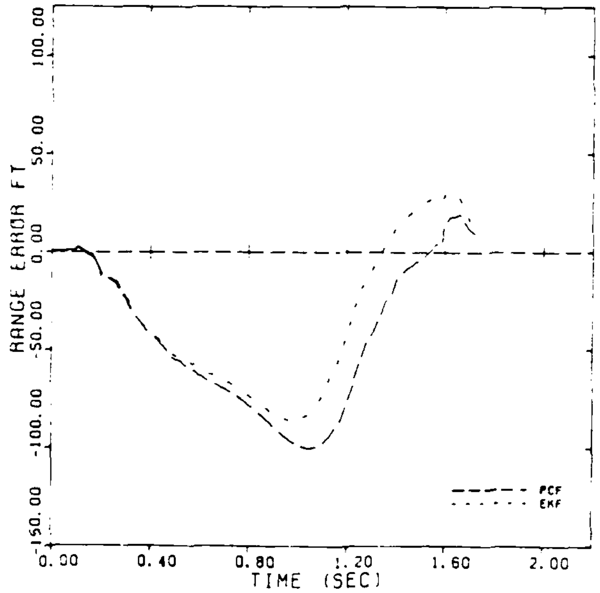


Fig. 5 Range error history (mismatch = 0.1).

The error histories of the PCF and the EKF for various conditions are given in Figs. 2-8. The error is defined as the difference between the magnitudes of the true and estimated range vectors. The magnitude of the estimated range is obtained by averaging the ten Monte Carlo runs. The estimated range in the PCF is computed as $(\hat{R}3)^{1/2}$.

The range-error histories for the nominal case are presented in Fig. 2 for the PCF and the EKF. The performances are quite similar. The range error history when the initial range has perturbations of 500 ft in the positive and negative directions are given in Figs. 3 and 4, respectively. The PCF clearly out-performs the EKF. Since both the observer and target are constantly maneuvering, the PCF, having states measurement functions, is able to utilize the information better than the EKF. By comparing the performances of the EKF in Figs. 3 and 4, it can be observed that the range errors, with a positive initial perturbation in range, are much worse than those with a negative initial perturbation. Thus the EKF is biased. The performance of the PCF, however, is more even with positive and negative perturbations.

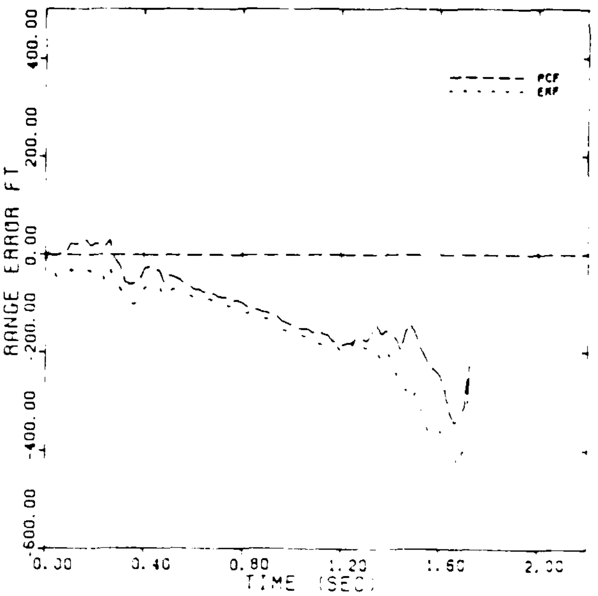


Fig. 6 Range error history (mismatch = 10).

Since the filter does not know the actual measurement noise variance, a measurement mismatch, defined as the ratio of the actual value to the assumed measurement covariance in the filters is hypothesized in generating the measurements. The range errors of the simulations with a measurement mismatch of 0.1 is presented in Fig. 5. The performance trends of both filters are similar to the nominal case in Fig. 2. For the higher mismatch of ten, both the PCF and the EKF are equally affected, as seen in Fig. 6.

To simulate actual situations, where uncertainties in both the states and the noise statistics can occur, experiments are made with perturbations to the initial states and measurement mismatches. The range error history with an initial state error of 500 ft and a measurement mismatch of 0.1 is presented in Fig. 7. With the same initial error and a measurement mismatch of ten the results in the range errors are given in Fig. 8. It is clear that in both cases, the PCF has a better response to perturbations than the EKF.

Biases Associated with Different Formulations of the Extended Kalman Filter

Observe from the given range error histories that the range estimate $\hat{R}_t$ of the EKF has a larger bias than the range estimate of the PCF. This phenomenon can be explained by developing and comparing the expressions for the square of the a posteriori range estimate of the EKF with that of the PCF. For simplicity, the differences in the expressions for the conditional range estimate are provided from a two-dimensional tracking problem. In this case, the polar coordinate for the PCF is (R^2, θ) where R is the range. The posterioric estimate of the range in the PCF is defined as $\hat{R}_p \overset{\Delta}{=} \hat{R}_p^2$, where $\hat{R}_p^2$ is the estimate of R^2. In order to illustrate the differences in the estimates of range with a different polar coordinate system, the expression for $\hat{R}_s$, the estimated range from the filter with a transformation from $\{x, y\}$ to $\{R, \theta\}$ is derived. It is proved that $\hat{R}_t \geq \hat{R}_p$ and shown that $\hat{R}_t \geq \hat{R}_s$. This inequality helps explain the smaller range biases of Mehra's polar coordinate filter[9] over the EKF for a reentry problem.

The Approximate Conditional Range from the EKF

It is assumeu that $x_o^T = (x_{R_o}, y_{R_o})$ is the state vector to be estimated with known s_o. The expressions for the a posteriori range estimate are developed at the first measurement where the known a priori conditions for all the filters are assumed to be the same. The state x_o has a Gaussian probability density function with mean $\bar{x}_o$ and covariance $\bar{P}_o$. The EKF for the estimation of x_o is given by

$$\hat{x}_o = \bar{x}_o + \Delta x_o \tag{21}$$

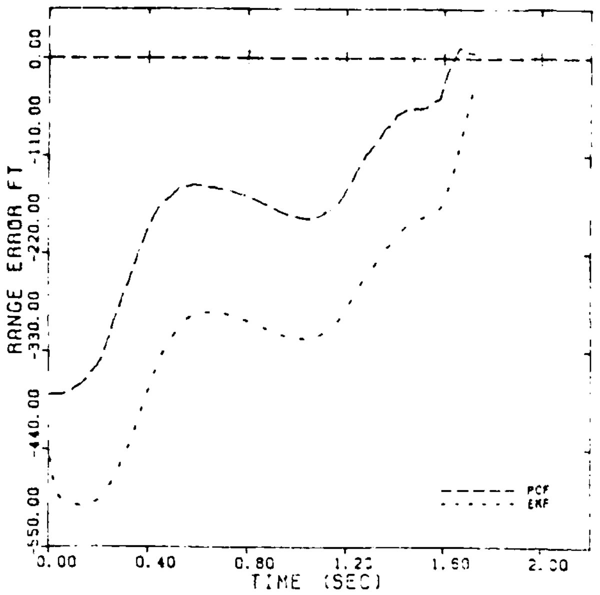


Fig. 7 Range error history (mismatch = 0.1, l = 500 ft).

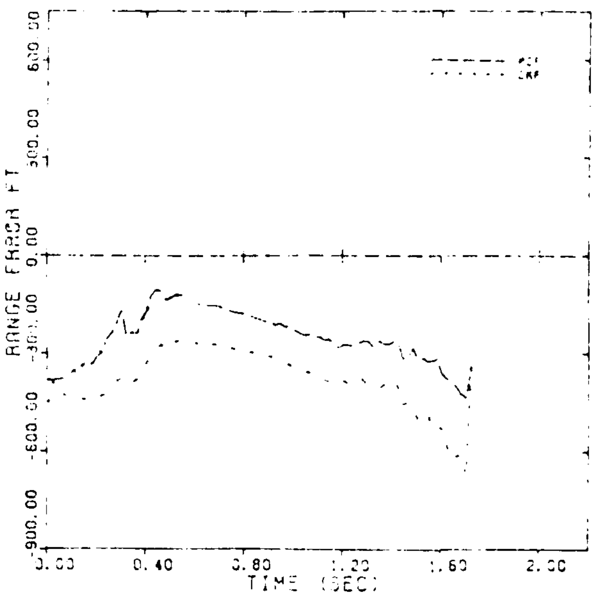


Fig. 8 Range error history (mismatch = 10, l = 500 ft).

where $\Delta x_o \triangleq \hat{P}_{x_o} \hat{H}_{x_o}^T V_o^{-1} \nu_o$, $\hat{P}_{x_o}$ is the a posteriori conditional covariance matrix; H_{x_o} is the first partial derivative matrix of the measurement function with respect to the state evaluated at $\bar{x}_o$; and ν_o equals $[z_o - h(\bar{x}_o)]$. The a posteriori estimate of the square of the range, $\hat{R}_{x_o}$, is computed from the EKF estimates

$$\hat{R}_{x_o}^2 \triangleq \hat{x}_o^T\hat{x}_o = \bar{x}_o^T\bar{x}_o + 2\bar{x}_o^T\Delta x_o + \Delta x_o^T\Delta x_o \tag{22}$$

The Conditional Range Estimates from the PCF

The transformation equations from the Cartesian system to a new polar system $y_o = [R_{p_o}^2, \theta_o]^T$ is given by $y_o = g(x_o)$ where $g(x_o) = [(x_{R_o}^2 + y_{R_o}^2), \tan^{-1}(y_{R_o}/x_{R_o})]^T$. This new polar coordinate system is chosen so that the Jacobian of the transformation is constant without approximations. Therefore, the conditional modes are invariant under this transformation.

The updates of the state $\hat{y}_o$ and the covariance $\hat{P}_{y_o}$ at the first measurement are calculated with Eqs. (18) and (19). The partial derivative of the measurement function in the y and x systems can be related at $H_{p_o} = (\partial\theta_o)/(\partial y_o) = H_{x_o}g_o^{-1}$. This expression for H_{p_o}, and Eq. (20) for $\hat{P}_{p_o}$ are substituted into Eq. (18) for $\hat{y}_o$ to yield

$$\hat{y}_o = \bar{y}_o + g_{x_o}^T\Delta x_o \tag{23}$$

The posterioric estimate of $\hat{R}_{p_o}^2$ is given by $(\bar{R}_{p_o}^2 = \bar{x}_o^T\bar{x}_o)$

$$\hat{R}_{p_o}^2 = \bar{R}_{p_o}^2 + 2\bar{x}_o^T\Delta x_o \tag{24}$$

The bias in the polar formulation can be demonstrated by differences in values of $\hat{R}_{p_o}^2$ from Eq. (24) and from $\hat{R}_{x_o}^2$ given by Eq. (22)

$$\hat{R}_{x_o}^2 - \hat{R}_{p_o}^2 = \Delta x_o^T\Delta x_o \tag{25}$$

This result is significant in the context of the numerical results of Figs. 3 and 4. The range error histories, in response to position perturbations in Figs. 3 and 4 show that the EKF seems positively biased. Since the analytical results show that $\hat{R}_{x_o}^2 > \hat{R}_{p_o}^2$, the relative performance of the PCF seems less biased than the EKF.

The Conditional Range Estimate of the Filter with R and θ as States

A formulation is now attempted along the same lines as Simmons[3] polar coordinate filter. The transformation of the rectangular states to a polar system $y_s \triangleq [R_{s_o}, \theta_o]^T$ where the measurements are linear in the states is given by $y_o = g(x_o)$ where

$$g(x_o) = [(x_{R_o}^2 + y_{R_o}^2)^{1/2}, \tan^{-1}(y_{R_o}/x_{R_o})]^T \tag{26}$$

The update equation for the range at a measurement is given by

$$\hat{R}_{s_o} = \bar{R}_{s_o} + \underline{g}_{1_o}\Delta x_o \tag{27}$$

where $\underline{g}_{1_o}$ contains the elements of g_o related to the $\hat{R}_{s_o}$ element, and is given by $\underline{g}_{1_o} = [\bar{x}_R/\bar{R}_{s_o}, \bar{y}_R/\bar{R}_{s_o}]$. The square of the estimated range computed by this filter is obtained for comparison with that of the PCF.

$$\hat{R}_{s_o}^2 = \bar{R}_{s_o}^2 + 2\bar{R}_{s_o}\underline{g}_{1_o}\Delta x_o + \Delta x_o^T\underline{g}_{1_o}^T\underline{g}_{1_o}\Delta x_o \tag{28}$$

where

$$\underline{g}_{1_o}^T\underline{g}_{1_o} = \begin{bmatrix} \bar{x}_R^2/\bar{R}_{s_o}^2 & \bar{x}_R\bar{y}_R/\bar{R}_{s_o}^2 \\ \bar{x}_R\bar{y}_R/\bar{R}_{s_o}^2 & \bar{y}_R^2/\bar{R}_{s_o}^2 \end{bmatrix} \tag{29}$$

Note that the determinant of the matrix $\underline{g}_{1_o}^T\underline{g}_{1_o}$ is zero, therefore, one eigenvalue is zero. Also, the trace of $\underline{g}_{1_o}^T\underline{g}_{1_o}$ is unity, and, therefore, the second eigenvalue is one. Consequently, $\underline{g}_{1_o}^T\underline{g}_{1_o}$ can be written as

$$\underline{g}_{1_o}^T\underline{g}_{1_o} = \Gamma_o \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Gamma_o^{-1} \tag{30}$$

where the columns of Γ_o are the eigenvectors of $\underline{g}_{1_o}^T\underline{g}_{1_o}$. The expression for $\hat{R}_{s_o}^2$ from Eq. (28) is rewritten with the expression for $\underline{g}_{1_o}^T\underline{g}_{1_o}$ from Eq. (30). By replacing $\bar{R}_{s_o}^2$ with $\bar{x}_o^T\bar{x}_o$,

$$\hat{R}_{s_o}^2 = \bar{x}_o^T\bar{x}_o + 2\bar{x}_o^T\Delta x_o + \Delta x_o^T\Gamma_o \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Gamma_o^{-1}\Delta x_o \tag{31}$$

By subtracting $\hat{R}_{s_o}^2$ from $\hat{R}_{p_o}^2$, given by Eq. (24), the differences in the range estimates obtained through the two different polar formulations can be shown

$$\hat{R}_{s_o}^2 - \hat{R}_{p_o}^2 = \Delta x_o^T\Gamma_o \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Gamma_o^{-1}\Delta x_o \tag{32}$$

Note that the right hand side of Eq. (32) is nonpositive. Therefore, $\hat{R}_c \geq \hat{R}_p$.

By differencing $\hat{R}_p^2$ in Eq. (31) from $\hat{R}_c^2$ (Eq. 22) an important result is obtained as

$$\hat{R}_{c_a}^2 - \hat{R}_{p_a}^2 = \Delta x_a^T \left[I - \Gamma_a \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \Gamma_a^{-1} \right] \Delta x_a$$

$$= \Delta x_a^T \Gamma_a \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Gamma_a^{-1} \Delta x_a \qquad (33)$$

The right hand side of Eq. (33) in this case is always non-negative. Therefore, the estimate of the range $\hat{R}_c$ with the Cartesian formulation of the EKF is usually greater than the range $\hat{R}_p$ obtained with an R and θ transformation for x and y.

Mehra[9] reported results from a reentry tracking problem where the range estimates of the EKF formulated in a rectangular Cartesian frame are always more positively biased than his polar coordinate filter. The inequality $\hat{R}_{c_a} \geqq \hat{R}_{p_a}$ helps to explain Mehra's results.

Conclusions

The problem of obtaining estimates of the relative states of a homing missile with respect to a target and the target acceleration, using bearings as the only measurements, has been considered. A new maximum likelihood filter has been formulated. This polar coordinate filter exploits the fact that the system dynamics are linear in rectangular coordinates and the measurements are linear in polar coordinates. The polar state space has been chosen so that the direct transformation of the maximum likelihood estimates between the spaces is preserved in two dimensions and, under certain assumptions, approximately in three dimensions. The numerical results show that the estimates of the range by the extended Kalman filter have a larger bias than that of the polar coordinate filter. Expressions for the posteriori estimates of the range for these filters have been derived and compared to corroborate the numerical results.

Acknowledgment

Research was partially supported by Eglin Air Force Base under contract F08635-82-C-0090.

References

[1] Fiske, P. H., "Advanced Digital Guidance and Control Concepts for Air-to-Air Tactical Missiles," Report AFATL-TR-77-130, Air Force Armament Laboratory, Armament Development and Test Center, Eglin Air Force Base, FL, Dec. 1977.

[2] Singer, R., "Estimating Optimal Tracking Filter Performance for Manned Maneuvering Targets," IEEE Transactions on Aerospace Electronics Systems, July 1970.

[3] Sammons, J. M., Balakrishnan, S., Speyer, J. L., and Hull, D. G., "Development and Comparison of Optimal Filters," Rep. AFATL-TR-79-87, Air Force Armament Laboratory, Air Force Systems Command, U.S. Air Force, Eglin Air Force Base, FL, Oct. 1979.

[4] Titus, H. (Ed.), "*Advances in Passive Target Tracking, Vol. I*," NPS-62TS77071, Naval Post Graduate School, Monterey, CA, May 1977.

[5] Tapley, B. D., Abusali, P. A. M., and Schutz, B. E., "Estimating the Motion of Maneuvering Targets Using Passive Measurements," IASM-TR-SO-2, Institute for Advanced Study in Orbital Mechanics, The University of Texas at Austin, Aug. 1980.

[6] Aidala, V. J. and Hammel, S. E., "Utilization of Modified Polar Coordinates for Bearings-Only Tracking," *IEEE Transactions on Automatic Control*, Vol. AC-28, No. 3, Mar. 1983.

[7] Tenney, R. R., Herbert, R. S., and Sandell, N. R., "A Tracking Filter for Maneuvering Sources," *IEEE Transactions on Automatic Control*, Vol. AC-22, No. 4, April 1977.

[8] Weiss, H. and Moore, J. B., "Improved Extended Kalman Filter Design for Passive Tracking," *IEEE Transactions on Automatic Control*, Vol. AC-25, No. 4, Aug. 1980.

[9] Mehra, R. K., "A Comparison of Several Nonlinear Filters for Reentry Vehicle Tracking," *IEEE Transactions on Automatic Control*, Vol. AC-6, No. 4, Aug. 1971.

[10] Jazwinski, A., "*Stochastic Processes and Filtering Theory*," Academic Press, New York, 1970.

[11] Smith, R. C., "A Six-Degrees of Freedom Computer Simulation of an Air-to-Air Missile Intercept," M.S. Thesis, The University of Texas at Austin, Texas, May 1979.

[12] Speyer, J. L. and Hull, D. G., "Comparison of Several Extended Kalman Filter Formulations for Homing Missile Guidance," *Proceedings of the AIAA Guidance and Control Conference*, Danvers, MA, Aug. 1980.

[13] Balakrishnan, S. N., "Development of Two Maximum Likelihood Estimation Methods and Their Applications to Tracking Problems," Ph.D. Dissertation, Dept. of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, May 1984.

On-Line Aircraft State and Stability Derivative Estimation Using the Modified-Gain Extended Kalman Filter

Jason L. Speyer* and Edwin Z. Cruest†
University of Texas, Austin, Texas

A new on-line state and parameter identification algorithm called the modified-gain extended Kalman filter (MGEKF) is applied to the problem of on-line state estimation and identification of the stability derivatives of a F-111 type of vehicle. The conceptual basis for the MGEKF is the existence of a class of nonlinear functions that allow a universal linearization with respect to the measurement function. This class includes the problem of identification of linear systems. The previous single-output formulation is extended to a multioutput formulation where the only available measurements are acceleration and pitch rate, but not elevator deflection. The filter formulation includes a simplified Dryden wind gust model. The inclusion of the wind gust model results mainly in a slowed response in the estimation of the stability derivatives associated with the acceleration state; estimates of the stability derivatives associated with the pitch rate still respond very quickly. The accuracy of the acceleration stability derivatives depends upon the amplitude and frequency components of the persistently exciting dither signal.

I. Introduction

THE historical development of aircraft parameter identification is given in Refs. 1-3. These studies are designed primarily for off-line use. A very complete study of recursive identification schemes for on-line use is given in Ref. 4. However, the usual assumption that the parameters be constant produces gains that are asymptotically inversely proportional to time and therefore become vanishing small. These schemes are not applicable to aircraft systems that must operate continuously and identify changes in the stability derivatives as the flight conditions change. One motivation for this type of on-line state and parameter estimation scheme is for use in adaptive flight control systems.

In Ref. 5, various schemes for identifying constant system parameters are compared on a common problem. Among these schemes is the extended Kalman filter (EKF) whose performance is shown to be relatively poor. This problem was again analyzed in Refs. 6 and 7, where a new estimation scheme called the modified-gain extended Kalman filter (MGEKF) is used For a special class of nonlinearities of which state and parameter estimation in linear systems is a member, there exists a universal linearization of these special nonlinearities with respect to the measurement function. In order to obtain nonlinearities in this class, the observability coordinate system rather than the controllability coordinate frame is used for the problem in Ref. 5. The results given in Ref. 7 indicate a remarkable improvement in performance. The MGEKF described in Ref. 7 is applied here to the problem of on-line state estimation and identification of the stability derivatives of an F-111 type of vehicle.

Section II presents the definition of a modifiable nonlinear system function that forms the basis of the MGEKF algorithm. A simple illustration of a modifiable nonlinearity is given before the general form of the MGEKF algorithm is stated. The essential features of this algorithm are then discussed. The dynamic system model for the aircraft is presented in Sec. III. The short-period longitudinal mode of the aircraft is expressed in acceleration and pitch rate states to be consistent with the measurements. In addition, a first-order model for the actuator and a second-order simplified Dryden wind gust model are described. In Sec. IV, the mechanization of MGEKF using this aircraft model is discussed and, in Sec. V, the performance of the MGEKF algorithm using accelerometer and pitch rate gyros is presented. Conclusions and recommendations are given in Sec. VI.

II. The Modified-Gain Extended Kalman Filter Algorithm

The dynamic nonlinear system model used for combined state and parameter estimation is presented first. The definition of a modifiable nonlinear system function, used as the basis of the MGEKF algorithm, is stated. Then the MGEKF algorithm is presented and its properties discussed.

Dynamical System and Modifiable Nonlinearities

The discrete dynamic system model used for combined state and parameters identification is

$$y_{i+1} = A(\theta_i)y_i + B(\theta_i)u_i + \underline{w}_i \tag{1}$$

$$\theta_{i+1} = \theta_i + \bar{w}_i \tag{2}$$

and the scalar measurement is

$$z_i = Hy_i + v_i = z_i^* + v_i \tag{3}$$

where y_i is an n-dimensional state vector, θ_i is a vector of maximal dimension $2n$ of unknown parameters representing the elements of the matrices $A(\theta_i)$ and $B(\theta_i)$, $A(\theta_i)$ is an $n \times n$ matrix, and $B(\theta_i)$ is an n vector where both contain up to n unknown elements represented by the elements of θ_i, u_i is a known scalar input, z_i^* is the scalar measurement function, H is a known $1 \times n$ measurement matrix, and $\underline{w}_i$, $\bar{w}_i$, and v_i are zero-mean white noise sequences with variances $\underline{Q}_i$, $\bar{Q}_i$, and γ_i, respectively. The formulation given here is for a single-input/single-output system consistent with the results of Ref. 2. Although the extension to more than one input is

Presented as Paper 85-1762 at the AIAA Atmospheric Flight Mechanics Conference, Snowmass, CO, Aug. 19-21, 1985; received Aug. 20, 1985; revision received July 24, 1986. Copyright © American Institute of Aeronautics and Astronautics, Inc., 1987. All rights reserved

*Harry H. Power Professor in Engineering, Department of Aerospace Engineering and Engineering Mechanics. Fellow AIAA.

†Research Assistant, Department of Aerospace Engineering and Engineering Mechanics. Member AIAA.

trivial, the extension to more than one output takes some innovation. This extension is done in the following sections.

The nonlinearity in this problem is $A(\theta_i)y_i$. For convenience define

$$x_i^T \triangleq [y_i^T, \theta_i^T] \tag{4}$$

and the nonlinearity as

$$f(x_i) \triangleq \begin{bmatrix} A(\theta_i)y_i + B(\theta_i)u_i \\ \theta_i \end{bmatrix} \tag{5}$$

where x_i has maximal dimension of $3n$.

Modifiable Nonlinearities

The notion of a modifiable nonlinearity is that there exists universal linearization of the function $f(x_i)$ with respect to the measurement function z_i^*.

Definition: A function f: $\boldsymbol{R}^p \rightarrow \boldsymbol{R}^p$ is a modifiable nonlinear system function if there exists a $p \times p$ matrix F: $\boldsymbol{R} \times \boldsymbol{R}^p \rightarrow \boldsymbol{R}^{p \times p}$ so that for any state x_i and known estimate of the state $\hat{x}_i$,

$$f(x_i) - f(\hat{x}_i) = F(z_i^*, \hat{x}_i)(x_i - \hat{x}_i) \tag{6}$$

where $z_i^* = Hx_i$.

Note that $F(z_i^*, \hat{x}_i)(x_i - \hat{x}_i)$ in Eq. (6) is a universal linearization of $f(x_i)$ with respect to the measurement function z_i^* without any approximation. Notice that the known function $F(z_i^*, \hat{x}_i) = F(Hx_i, \hat{x}_i) \neq F(H\hat{x}_i, \hat{x}_i)$, where the latter quantity is the differential of f evaluated at $\hat{x}$, as used in the linearization.

The noiseless case of a simple linear dynamical system with an unknown coefficient illustrates the idea of a modifiable nonlinearity. The system is represented as

$$y_{i+1} \triangleq \theta_i y_i, \qquad \theta_{i+1} = \theta_i, \qquad z_i^* = y_i \tag{7}$$

where y_i is a scalar state at stage i and θ_i the unknown parameter at stage i. The nonlinearity is put into modifiable form by writing

$$\begin{aligned} (x_{i+1} - \hat{x}_{i+1}) &= \begin{bmatrix} \theta_i y_i \\ \theta_i \end{bmatrix} - \begin{bmatrix} \hat{\theta}_i \hat{y}_i \\ \hat{\theta}_i \end{bmatrix} \\ &= \begin{bmatrix} \theta_i \hat{y}_i - \hat{\theta}_i y_i + \hat{\theta}_i y_i - \hat{\theta}_i \hat{y}_i \\ \theta_i - \hat{\theta}_i \end{bmatrix} \\ &= F(z_i^*, \hat{x}_i)(x_i - \hat{x}_i) \end{aligned} \tag{8}$$

where $x_i^T \triangleq [y_i, \theta_i]$, $\hat{y}$ and $\hat{\theta}$ are the estimated values of y and θ, and

$$F(z_i^*, \hat{x}_i) = \begin{bmatrix} \hat{\theta}_i & z_i^* \\ 0 & 1 \end{bmatrix}, (x_i - \hat{x}_i) = \begin{bmatrix} y_i - \hat{y}_i \\ \theta_i - \hat{\theta}_i \end{bmatrix} \tag{9}$$

Note that the estimation error in Eq. (8) is propagated without approximations by a linear equation. Since the measurements are linear, the update formula for the error is also linear. In the noiseless case where this filter reduces to a nonlinear observer, the error of this observer is shown to be exponentially convergent by Lyapunov's second method.[6,7] In the noise-corrupted case where only the noisy measurement is available and not the measurement function, under certain a priori uncheckable conditions the MGEKF is shown to be exponentially bounded in the mean square sense.[6,7]

The MGEKF Algorithm

The discrete formulation of the MGEKF from Ref. 7, based on the dynamic system [Eqs. (1-3)] using Eqs. (4) and (5), is summarized as

$$\bar{x}_{i+1} = f(\hat{x}_i) \tag{10}$$

$$\hat{x}_i = \bar{x}_i + K_i(z_i - H\bar{x}_i) \tag{11}$$

$$K_i = M_i H^T (HM_i H^T + \gamma_i)^{-1} \tag{12}$$

$$M_{i+1} = F(z_i, \hat{x}_i) P_i F(z_i, \hat{x}_i) + Q_i \tag{13}$$

$$P_i = (I - K_i H) M_i (I - K_i H)^T + K_i \gamma_i K_i^T \tag{14}$$

where $\bar{x}_i$ is the propagated state and parameter estimates, K_i the modified Kalman gain calculated by Eqs. (12-14), M_{i+1} the propagated pseudo error-covariance matrix, P_i the updated pseudo error-covariance matrix, Q_i the process noise covariance matrix composed of diagonal matrix elements $\underline{Q}_i$ and $\overline{Q}_i$, and γ_i the measurement noise covariance matrix.

The functions $f(\hat{x}_i)$ and $F(z_i, \hat{x}_i)$ can be expressed in a simple way when the dimension of θ_i is assumed to be $2n$. Note that $B(\theta_i)$ becomes just the last n elements of θ. Furthermore, the function $F(z^*, \hat{x}_i)$ obtained from a modifiable nonlinear function $f(x_i)$ becomes

$$F(z_i^*, \hat{x}_i) = \left[\begin{array}{c:c:c} A(\hat{\theta}_i) & z_i^* I_n & u_i I_n \\ \hdashline 0_{2n \times n} & \multicolumn{2}{c}{I_{2n}} \end{array}\right] \tag{15}$$

where I_n is an $n \times n$ identity matrix and $0_{2n \times n}$ is a $2n \times n$ matrix of zeros. In Ref. 7, this matrix is obtained in the observability canonical form where the unknown parameters lie in the last column of the A matrix. It should be noted that in the gain algorithm for propagating the pseudo error-covariance matrix [Eq. (13)], the actual measurement z_i is used rather than the measurement function z^* in Eq. (15). Finally, for use later when describing the MGEKF for the aircraft application

$$f(\hat{x}_i) = F(0, \hat{x}_i)\hat{x}_i \tag{16}$$

The key to applying the MGEKF to the parameter estimation problem is to ensure that those unknown parameters being identified enter the dynamic equation so as to multiply the states or controls that are directly measured. As shown in Ref. 7, this means that the coordinate frame must be chosen carefully. Furthermore, the results given in Ref. 7 apply to only a single output problem. The results here give an example of how the MGEKF can be extended to two or more outputs.

III. The Aircraft Dynamical System

The linear longitudinal dynamics representing the short period motion are

$$\dot{\alpha} = Z_\alpha \alpha + Z_q q + Z_e e - Z_\alpha \alpha_G + b_\alpha \tag{17}$$

$$\dot{q} = M_\alpha \alpha + M_q q + M_e e - M_\alpha \alpha_G + b_q \tag{18}$$

where α is the total angle of attack, q the pitch rate, e the elevator deflection, α_G the angle of attack due to wind gust, b_α and b_q the trim biases associated with the steady-state conditions of α and q, respectively, and Z_α, Z_q, Z_e, M_α, M_q, and M_e the aircraft stability derivatives.

Transformation of State Space

Aircraft, such as that in Ref. 8 and the F-111 type, have normal acceleration and pitch rate measurements available from an accelerometer and pitch rate gyro. Therefore, it is advantageous to convert from angle of attack to acceleration in the dynamical representation of the aircraft for MGEKF applications.

The accelerometer measures the combined acceleration of the center of mass and the acceleration relative to the center of mass due to the moment arm X_{acc}. This relationship between A, q, and α is

$$A = \beta_1(q - \dot{\alpha}) + \beta_2\dot{q} \qquad (19)$$

where A is the normal acceleration (in g's) and

$$\beta_1 \triangleq \frac{2\pi u}{360g}, \qquad \beta_2 \triangleq \frac{2\pi X_{acc}}{360g} \qquad (20)$$

where u is the aircraft velocity (in ft/s), g the gravitational acceleration, and X_{acc} the X distance from the aircraft c.g. to the accelerometer.

Equations (17-19) can be used to derive a new set of dynamical equations using A and q.[8] The new dynamical system is

$$\dot{A} = D_A A + D_q q + D_e e + D_G \alpha_G + \Omega_e \dot{e} + \Omega_G \dot{\alpha}_G + B_A \quad (21)$$

$$\dot{q} = H_A A + H_q q + H_e e + H_G \alpha_G + B_q \qquad (22)$$

where

$$D_A \triangleq Z_\alpha + \frac{M_\alpha \Omega_q}{\Omega_\alpha}$$

$$D_s = Z_s\Omega_\alpha - Z_\alpha\Omega_s + M_s\Omega_q - \frac{M_\alpha\Omega_q\Omega_s}{\Omega_\alpha}, \qquad s = q, e, G$$

$$H_A \triangleq \frac{M_\alpha}{\Omega_\alpha}; \qquad H_s \triangleq M_s - \frac{M_\alpha\Omega_s}{\Omega_\alpha}, \qquad s = q, e, G$$

$$\Omega_s \triangleq \beta_2 M_s - \beta_1 Z_s, \qquad s = \alpha, e, G$$

$$\Omega_q \triangleq \beta_2 M_q + \beta_1(1 - Z_q)$$

$$B_A = \Omega_q - \frac{\Omega_\alpha\beta_2}{\beta_1}B_q, \qquad B_q \triangleq \frac{\beta_1 b_\alpha - \beta_2 b_q}{\Omega_\alpha} + b_q$$

Note that Z_G and M_G are the stability derivatives for acceleration and pitch rate associated with wind gust effects. It can be seen from Eqs. (17) and (18) that $Z_G = -Z_\alpha$ and $M_G = -M_\alpha$; if these relationships hold, they imply that $D_G = 0$ and $H_G = 0$. Interestingly, this results in only the acceleration equation being directly affected by the wind gusts and then only by $\dot{\alpha}_G$. Although not directly affected by wind gusts, the pitch rate is affected by the wind gusts through the acceleration term. The aircraft dynamics can now be written as

$$\dot{A} = D_A A + D_q q + D_e e + \Omega_e \dot{e} - \Omega_\alpha \dot{\alpha}_G + B_A \qquad (23)$$

$$\dot{q} = H_A A + H_q q + H_e e + B_q \qquad (24)$$

Note that B_A and B_q are biases associated with A and q, respectively. In the next sections, the elevator actuator and wind gust models are described.

Elevator Dynamics

A measurement of the elevator deflection is not available. The first-order elevator actuator dynamics, which determine the actual position of the elevator in response to an elevator command e_c, are assumed to be of the form

$$\dot{e} = -H_1 e + H_2 e_c \qquad (25)$$

where H_1 and H_2 reflect the dominant dynamical characteristic of the elevator actuator. As in Ref. 8, the actuator dynamic coefficients are assumed to remain *constant*; therefore, H_1 and H_2 need *not* be estimated. However, since the actual elevator deflection is not measured, it must be estimated on-line.

Wind Gust Dynamics

The wind gust model is also included in the formulation. Two commonly accepted wind gust models used in the analysis of aircraft motion are the Dryden and the von Kármán models.[9] For estimation purposes, a simplified Dryden model[9,10] which compares very well with the Dryden model, is used. The simplified Dryden wind gust model is

$$\ddot{\alpha}_G = K_1\alpha_G + K_2\dot{\alpha}_G + \omega_G(t) \qquad (26)$$

where ω_G is a zero-mean white noise process with spectral density $K_3^2 Q_w$ and

$$Q_w = 2\frac{C_2}{C_3}\frac{\sigma^2 L_w}{u} \qquad C_2 = \frac{1 + 3\beta}{2\beta^{4/3}}, \qquad C_3 = \frac{(1 + \beta)^{4/3}}{\beta^{4/3}}$$

$$\beta = \frac{b}{2L_w}, \qquad L_w = L_\infty \frac{h}{h + h_0}$$

$$K_1 = -C_3\frac{u^2}{L_w^2}, \qquad K_2 = -C_2\frac{u}{L_w}, \qquad K_3 = C_3\frac{u}{L_w^2}$$

where h is the altitude (in ft), $h_0 = 2500$ ft, $L_\infty = 2000$ ft, b the wing span (in ft), u the aircraft velocity (in ft/s), and σ the rms gust velocity (in ft/s). By letting $\alpha_1 = \alpha_G$ and $\alpha_2 = \dot{\alpha}_G$, the following set of linear equations is obtained from Eq. (26):

$$\dot{\alpha}_1 = \alpha_2 \qquad (27)$$

$$\dot{\alpha}_2 = K_1\alpha_1 + K_2\alpha_2 + K_3\omega_G(t) \qquad (28)$$

Augmented State Variables and Modified Nonlinearities

If Eqs. (27), (28), and (25) are augmented to the dynamics of Eqs. (23) and (24), then the dynamical system is expanded to fifth order in the states A, q, e, α_1, and α_2. If in addition, the constant biases B_A and B_q are augmented as states, then we obtain a seventh-order system.

However, the system dynamics are *not* modifiable since the unmeasured state e multiplies some of the parameters to be estimated. However, the system dynamics can be made modifiable by replacing state e with *two* new states defined as

$$X_1 \triangleq \bar{D}_e e, \quad X_2 \triangleq H_e e \qquad (29)$$

where $\bar{D}_e \triangleq D_e - H_1\Omega_e$. Note that e_c, the commanded elevator position, is available and is assumed to be known perfectly. This results in the following modifiable dynamical system:

$$\begin{bmatrix} \dot{A} \\ \dot{q} \\ \dot{X}_1 \\ \dot{X}_2 \\ \dot{\alpha}_1 \\ \dot{\alpha}_2 \\ \dot{B}_A \\ \dot{B}_q \end{bmatrix} = \begin{bmatrix} D_A & D_q & 1 & 0 & 0 & -\Omega_\alpha & 1 & 0 \\ H_A & H_q & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & -H_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -H_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & K_1 & K_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\times \begin{bmatrix} A \\ q \\ X_1 \\ X_2 \\ \alpha_1 \\ \alpha_2 \\ B_A \\ B_q \end{bmatrix} + \begin{bmatrix} \Omega_e \\ 0 \\ H_2\bar{D}_e \\ H_2 H_e \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} e_c + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \omega_G \\ 0 \\ 0 \end{bmatrix} \qquad (30)$$

where $\bar{\Omega}_e \triangleq H_2\Omega_e$. Note that by this device the parameters $\bar{D}_e$ and H_e are now multiplied by e, which is a known input. Thus, with acceleration and pitch rate measurements, the nonlinearities are modifiable nonlinearities. As stated in Ref. 7, it is important that the system be observable. By using the observability test in Ref. 11, it can be shown that this system is observable.

Transformation from Continuous to Discrete Form

Since the discrete version of the MGEKF will be implemented, the aircraft dynamical equations must be transformed into discrete form. By assuming the sample time Δt to be sufficiently small, the discrete dynamical equations are approximately

$$\begin{bmatrix} A \\ q \\ X_1 \\ X_2 \\ \alpha_1 \\ \alpha_2 \\ B_A \\ B_q \end{bmatrix}_{i+1} = \begin{bmatrix} D_A & D_q & 1 & 0 & 0 & R_3 & \Delta t & 0 \\ H_A & H_q & 0 & 1 & 0 & 0 & 0 & \Delta t \\ 0 & 0 & C_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & C_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \Delta t & 0 & 0 \\ 0 & 0 & 0 & 0 & R_1 & R_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}_i$$

$$\times \begin{bmatrix} A \\ q \\ X_1 \\ X_2 \\ \alpha_1 \\ \alpha_2 \\ B_A \\ B_q \end{bmatrix}_i + \begin{bmatrix} S_A \\ S_q \\ C_2 D_e \\ C_2 H_e \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_i e_{c_i} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \omega_{a1} \\ \omega_{a2} \\ 0 \\ 0 \end{bmatrix} \qquad (31)$$

$$D_{A_i} \approx D_A \Delta t + 1, \qquad D_{q_i} \approx D_q \Delta t$$

$$D_{e_i} \approx \bar{D}_e \Delta t, \qquad H_{A_i} \approx H_A \Delta t$$

$$H_{q_i} \approx H_q \Delta t + 1, \qquad H_{e_i} \approx H_e \Delta t$$

$$S_{A_i} \approx H_2 [\Omega_e (D_{A_i} + 1) + D_{e_i}] \frac{\Delta t}{2}$$

$$S_{q_i} \approx H_2 (\Omega_e H_{A_i} + H_{e_i}) \frac{\Delta t}{2}$$

Note that $x_{1_i} = \bar{D}_e e_i$ and $x_{2_i} = H_e e_i$ and that the following parameters are assumed known at each time step i:

$$C_1 = e^{-H_1 \Delta t}, \qquad C_2 = \frac{H_2}{H_1}(1 - C_1)$$

$$\omega_{a1} = R_4 \eta_G, \qquad \omega_{a2} = R_5 \eta_G$$

$$R_1 \approx K_1 \Delta t, \qquad R_2 \approx K_2 \Delta t + 1, \qquad R_3 \approx -\Omega_\alpha \Delta t$$

$$R_4 \approx \frac{K_3 (\Delta t)^2}{2}, \qquad R_5 \approx \frac{(2 + K_2 \Delta t) K_3}{2}$$

where η_G is a zero-mean noise sequence with variance $Q_w / \Delta t$. In making the discrete approximation, the exact discrete form is used when convenient; otherwise, the above is the first term of a Taylor series in Δt.

IV. Implementing the MGEKF

The MGEKF algorithm, formulated in Eqs. (10-14), is to be applied to the problem of estimating the aircraft states and parameters. This algorithm is extended from a single output to the output of acceleration and pitch rate. We have already shown that the dynamic nonlinearities are modifiable nonlinearities. The acceleration and pitch rate measurements are

$$z_{A_i} = A_i + v_{A_i}, \qquad z_{q_i} = q_i + v_{q_i} \qquad (32)$$

where v_A and v_q are constant standard deviations associated with the measurement noise of A and q, respectively, so that

$$\gamma_i = \begin{bmatrix} v_A^2 & 0 \\ 0 & v_q^2 \end{bmatrix} \qquad (33)$$

The formulation of the MGEKF algorithm requires that the matrices $F(z_i, \hat{x}_i)$, H, and $f(\hat{x}_i)$ be formed from the aircraft dynamic system [Eq. (31)] and measurements [Eq. (32)]. Since we did not include the biases (B_A, B_q) in our linear simulation, they are not included in the state space defined now as

$$x_i^T \triangleq [A, q, X_1, X_2, \alpha_1, \alpha_2, D_A, D_q, H_A, H_q, S_A, S_q, D_e, H_e]_i \qquad (34)$$

and the measurement is defined as the two-vector

$$z_i \triangleq [z_A, z_q]_i^T \qquad (35)$$

Then

$$H = [I_{2\times 2}, 0_{2\times 12}] \qquad (36)$$

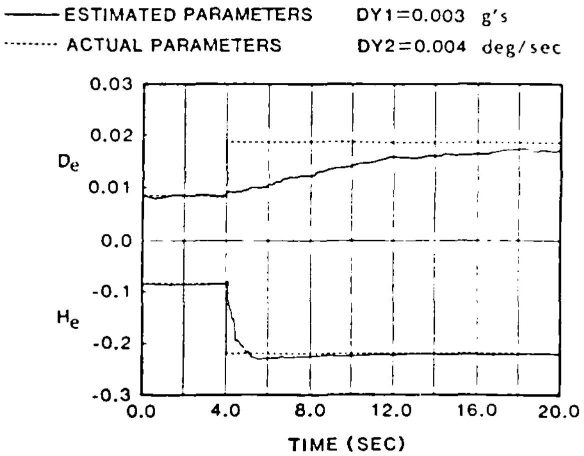


Fig. 1 Parameter tracking with a step change in flight conditions. WG = 1 ft/s, accurate instruments, and low-amplitude dither signal.

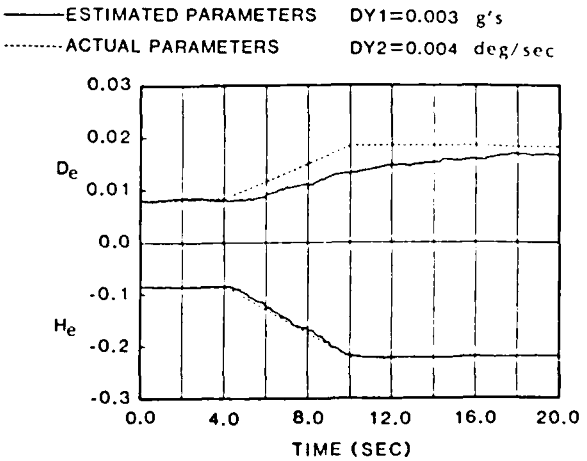


Fig. 2 Parameter tracking with a ramp change in flight conditions. WG = 1 ft/s, accurate instruments, and low-amplitude dither signal.

$$F(z_i,\hat{x}_i)=\left[\begin{array}{cccccccccccccc}\hat{D}_A & \hat{D}_q & 1 & 0 & 0 & R_3 & z_A & z_q & 0 & 0 & e_c & 0 & 0 & 0\\ \hat{H}_A & \hat{H}_q & 0 & 1 & 0 & 0 & 0 & 0 & z_A & z_q & 0 & e_c & 0 & 0\\ 0 & 0 & C_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & C_2e_c & 0\\ 0 & 0 & 0 & C_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & C_2e_c\\ 0 & 0 & 0 & 0 & 1 & \Delta t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & R_1 & R_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ & & 0_{8\times 6} & & & & & & I_{8\times 8} & & & & & \end{array}\right]_i \tag{37}$$

$$f(\hat{x}_i)=F(0,\hat{x}_i)\hat{x}_i \tag{38}$$

Note the placement of the measurements in the error dynamics matrix $F(z_i,\hat{x}_i)$ [see Eqs. (6) and (13)]. Each measurement multiplies the corresponding parameter error for that particular state. For example, the acceleration measurement z_A multiplies its corresponding parameter error $(D_A-\hat{D}_A)$. No parameter associated with any of the states except acceleration and pitch rate is identified. Since input e_c is multiplied by C_2 in the rows associated with states X_1 and X_2, the estimated parameters are now D_e and H_e rather than C_2D_e and C_2H_e.

The Process Noise Covariance Matrix and Filter Tuning

Process noise is assumed for the parameters [see Eq. (2)] in order to keep the filter gains associated with the parameters from going to zero. The values of the process noise variance for the parameters are chosen by tuning the filter to obtain best performance. The discrete stochastic equation is

$$x_{i+1}=F(0,x_i)x_i+w_i \tag{39}$$

where w_i is composed of the six-dimensional vector $\underline{w}_i$ associated with the states and the eight-dimensional vector $\overline{w}_i$ associated with the parameters. Note that $w_3(i)$, $w_4(i)$, $w_5(i)$, and $w_6(i)$ are elements of $\overline{w}_i$ that have the state-dependent forms

$$w_3(i)=(C_1e+C_2e_c)_iw_{13}(i)+D_ew_e(i)$$

$$w_4(i)=(C_2e+C_2e_c)_iw_{14}(i)+H_ew_e(i)$$

$$w_5(i)=\omega_{a1}(i),\qquad w_6(i)=\omega_{a2}(i)$$

while the remaining noise processes are assumed independent.

The process noise covariance matrix $Q_i=E[w_iw_i^T]$ has values along its diagonal and the only nonzero off-diagonal elements are

$$Q_{3,4}=Q_{4,3}=D_eH_eQ_{3,3}$$

$$Q_{5,5}=Q_{6,5}=R_4R_5Q_w/\Delta t$$

$$Q_{3,13}=Q_{13,3}=\bar{e}Q_{13,13}$$

$$Q_{4,14}=Q_{14,4}=\bar{e}Q_{14,14}$$

$$\bar{e}=C_1e+C_2e_c$$

where the dependence on the time stage i has been suppressed. Since some of the off-diagonal elements of Q_i depend on the parameters or states, they must be approximated. The adaptive form of the process noise covariance matrix suggested in Ref. 7 uses the current estimates of the parameters to adapt Q_i.

The first choice for the process noise standard deviations associated with the states other than wind gusts was to set them to zero, since no modeled process noises exist on any of the states except wind gusts. However, to enhance the MGEKF performance in the presence of modeling inaccuracies, small power spectral densities are assumed. Although the same reasoning can be used in choosing the standard deviations associated with the estimated parameters, it was discovered experimentally that setting the standard deviation of the parameters to about the magnitude of the difference between the minimum and maximum values of the parameters over the flight envelope yielded the best results.

V. Computer Simulation and Control System Design Consideration

The aircraft simulation routine provides the MGEKF routine with acceleration A, pitch rate q, commanded elevator position e_c, forward velocity u, and altitude h. The MGEKF uses the measurements of A, q, and e_c directly, while the wind gust model uses u and h to calculate wind gust coefficients in $F(z_i,\hat{x}_i)$.

The simulation uses an exact discrete form of the continuous dynamic equations. The trim biases are not included. The dynamic system is persistently excited by an oscillatory dither of the elevator input in order to estimate the parameters. The dither signal maintains an adequate signal-to-noise ratio in the absence of pilot input, enabling the filter to differentiate between the response of the dynamical system and noise on the system.[4]

Shape, amplitude, and frequency are the three major aspects of the dither signal important to the performance of the MGEKF. A sinusoidal dither signal composed of three frequencies gave good results. Two frequencies are at the high (4.3 rad/s) and low (1.8 rad/s) ends of the expected short-period frequency of the aircraft over the flight envelope. The third is at the frequency of the higher-order actuator (20 rad/s) because experimentation indicated that the parameters associated with the control (S_A, S_q, D_e, and H_e) are more easily identified if a frequency corresponding to the natural frequency of the actuator is included in the dither. Use of these frequencies results in the improved performance of the MGEKF, which allows a decrease in the amplitude of the dither signal while still maintaining performance.

In order to determine the effect of sensor accuracy on the performance of the filter, the accelerometer noise standard deviations are alternately set to 0.003 and 0.03 g. The pitch rate gyro noise standard deviations are alternately set to 0.004 and 0.04 deg/s. Finally, three levels of clear air turbulence were considered: $\sigma=$ WG = 1, 5, and 15 ft/s.

The objective of our numerical experiments are to show how the MGEKF tracks the states and stability deviations through a change in flight conditions from an altitude of 15,000 ft and a Mach number of 0.6 to an altitude of 500 ft and a Mach number of 0.69. In the linear simulation, the transition from one flight condition to the other was formed either as a step or a ramp. The ramp is essentially a linear interpolation between the parameters associated with each flight condition over a 6 s period. Since only the dither signal excites the aircraft during the transition, it is expected that the performance shown here is somewhat conservative, since changes in flight conditions require control inputs that will generate additional acceleration and pitch rate.

The final conditions of a 10 s convergence run are used as the initial conditions for each run. This represents a *steady-state* starting condition. A ramp change from one aircraft flight condition to another occurs between 4.0 and 10.0 s, with no changes made from 10.0 s to the end of the run at 20.0 s.

In the following figures, solid lines represent the estimated parameter and the dashed lines represent the actual parameter. Also printed on these figures are DY1 and DY2, the standard deviations of the measurement noise on the accelerometer and the pitch rate gyro, respectively. WG is the rms value of the wind gusts, which indicates the process noise on the system. The sample frequency is 100 Hz.

Only the parameters D_e and H_e are used to compare results of variations in system design and model design, because D_e is a good indicator of the tracking characteristics of the other parameters associated with the acceleration equations and H_e is a good indicator of the pitch rate parameters.

A step jump in flight conditions shown in Fig. 1 indicates the step response characteristics of the MGEKF. The slow response in $\hat{D}_e$ is characteristic of the parameters associated with the acceleration state. This is due to the wind acceleration term in the acceleration equation. In contrast, note the rapid response of the estimated parameter $\hat{H}_e$.

The Dryden wind gust model is obtained empirically from many atmospheric studies[10]; therefore, the wind gust characteristics of the real atmosphere will not exactly correspond to the assumed wind gust model in Eqs. (22) and (28). This fact must be considered when analyzing the performance of the MGEKF, since its performance may suffer if the actual

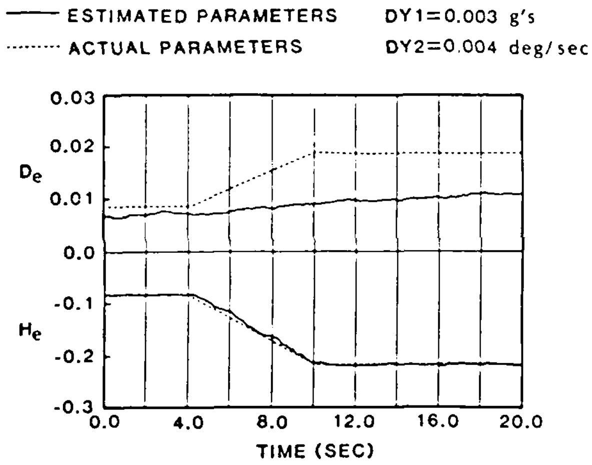


Fig. 3 Parameter tracking with a ramp change in flight conditions, WG = 5 ft/s, accurate instruments, and low-amplitude dither signal.

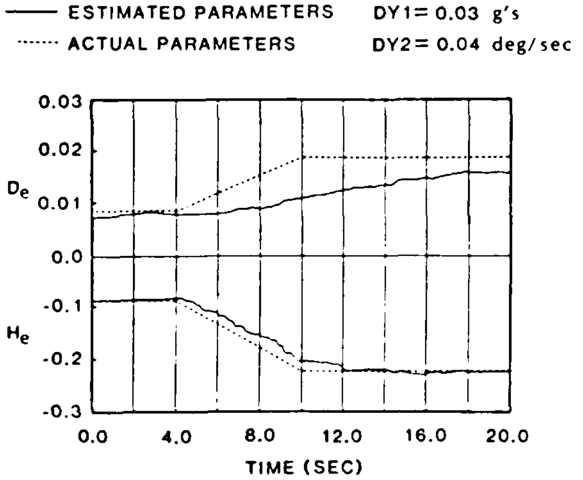


Fig. 4 Parameter tracking with a ramp change in flight conditions, WG = 1 ft/s, reduced accuracy instruments, and low-amplitude dither signal.

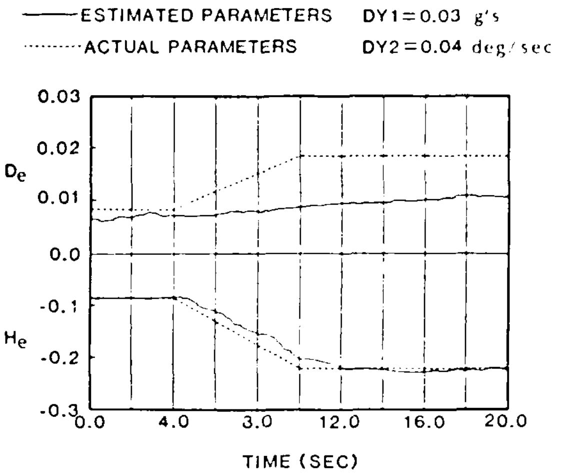


Fig. 5 Parameter tracking with a ramp change in flight conditions, WG = 5 ft/s, reduced accuracy instruments, and low-amplitude dither signal.

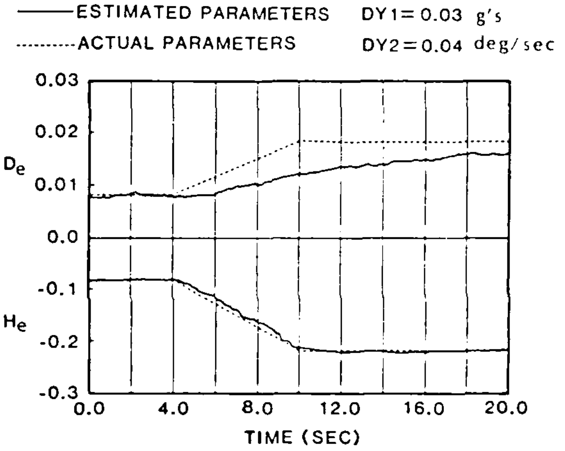


Fig. 6 Parameter tracking with a ramp change in flight conditions, WG = 5 ft/s, reduced accuracy instruments, and increased-amplitude dither signal.

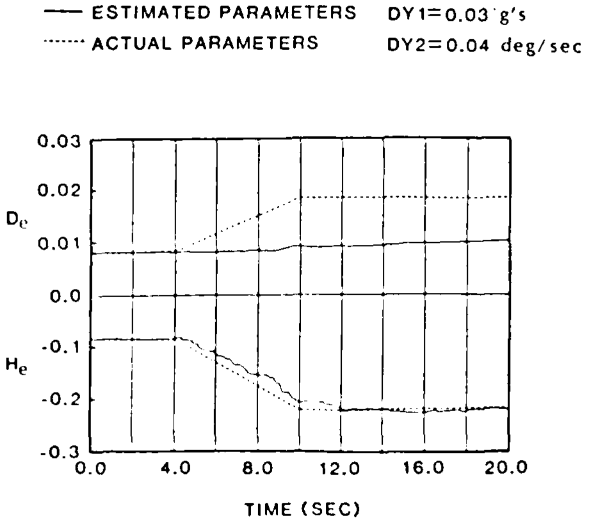


Fig. 7 Parameter tracking with a ramp change in flight conditions, WG = 5 ft/s in filter, WG = 1 ft/s in simulation, reduced accuracy instruments, and low-amplitude dither signal.

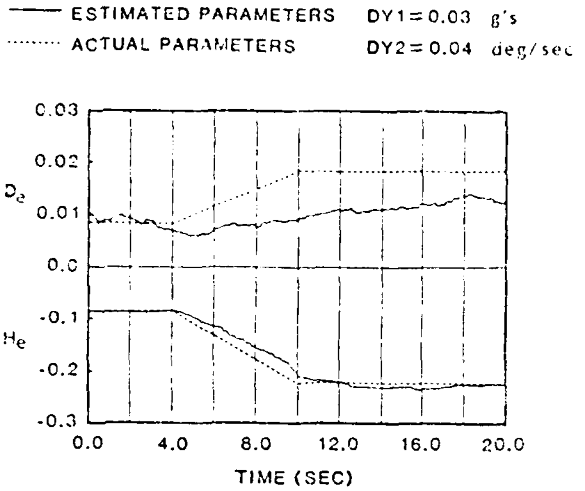


Fig. 8 Parameter tracking with a ramp change in flight conditions, WG = 5 ft/s in filter, WG = 15 ft/s in simulation, reduced-accuracy instruments, and low-amplitude dither signal.

wind gust characteristics are mismatched. The effects are negligible at WG = 5 ft/s when a run having a 15,000 ft mismatch between the actual wind gust dynamics and the wind gust model is compared with the matched case with all other factors being the same.

For the same dither signal and instrument accuracy, Figs. 2 and 3 show the effect of changing clear air turbulene from WG = 1 to 5 ft/s, where the corresponding rms acceleration over the 20-s flight for each profile is 0.0926 and 0.1195 g, respectively. The effect of reduced instrument accuracy on the low-amplitude dither signal is shown in Figs. 4 and 5. Note that, with clear air turbulance of WG = 5 ft/s, comparing Figs. 3 and 5 shows that the effect of an improved measurement device has little effect on D_e but has some effect on H_e. The effect of an increase in dither signal amplitude is shown in Fig. 6, where the rms acceleration over the 20-s flight is 0.4 g. Note that now Figs. 2 and 6 have about the same profile even with reduced instrument accuracy. However, large accelerations and pitch rates produce pilot and passenger discomfort. Finally, Figs. 7 and 8 show the mismatched performance of the filter of Fig. 5 (WG = 5 ft/s) with clear air turbulence in the simulation is lower (WG = 1 ft/s) and higher (WG = 15 ft/s) than WG = 5 ft/s, respectively. Note that R_3 in Eq. (31), which can be determined from the estimated parameters, need be chosen only at some nominal value since WG is never known. That is, if $R_3\alpha_2$ is chosen as a state variable, then the wind intensity becomes R_3WG. The results shown are for a Monte Carlo average of 10 runs.

VI. Conclusion

Although additional investigation of the modified-gain extended Kalman filter (MGEKF) is needed, the results from this study indicate that the MGEKF displays reasonable performance in the areas of convergence characteristics, disturbance rejection, and response to system changes. The slow response of the parameter estimation error associated with the acceleration equation is due mostly to the inclusion of the high-frequency gust acceleration term. Improvements might be made by adding an angle-of-attack meter that measures the relative motion between the aircraft and the air mass. Therefore, the MGEKF seems well suited to application in an adaptive control scheme. The MGEKF can provide state and parameter estimates to a set of control laws that use these estimates to adapt to changing flight conditions. The controller should be designed to rely most heavily on the parameters associated with pitch rate. This is not unreasonable since H_e, which is essentially the change in moment due to elevator deflection, is estimated well and is important in designing responsive flight control systems.

Acknowledgment

This work was sponsored by General Dynamics, Fort Worth Division, and the U.S. Air Force Office of Scientific Research under Grant AFOSR-84-0371.

References

[1] Denery, D.G., "Identification of System Parameters From Input-Output Data With Applications to Air Vehicles," NASA TN D-6468, Aug. 1971.

[2] Rediess, H.A., "An Overview of Parameter Estimation Techniques and Applications in Aircraft Flight Testing," *Proceedings of Symposium on Parameter Estimation Techniques and Applications*, NASA TN D-7647, April 1974, pp. 1-18.

[3] Iliff, K.W. and Maine, R.E., "Practical Aspects of Using a Maximum Likelihood Estimation Method to Extract Stability and Control Derivatives From Flight Data," NASA TN D-8209, 1976

[4] Ljung, L. and Soderstrom, T., *Theory and Practice of Recursive Identification*, Massachusetts Institute of Technology Press, Cambridge, 1983.

[5] Saridis, G.H., "Comparison of Six On-Line Identification Algorithms," *Automatica*, Vol. 10, 1974.

[6] Song, T.L. and Speyer, J.L., "A Stochastic Analysis of a Modified Gain Extended Kalman Filter with Applications to Estimation with Bearings Only Measurements," *IEEE Transactions on Automatic Control*, Vol. AC-30, Oct. 1985.

[7] Song, T.L. and Speyer, J.L., "The Modified Gain Extended Kalman Filter and Parameter Identification in Linear Systems," *Automatica*, Vol. 22, No. 1, Jan. 1986, pp. 59–75.

[8] Speyer, J.L., White, J.E., Douglas, R., and Hull, D.G., "Multi-Input/Multi-Output Controller Design for Longitudinal Decoupled Aircraft Motion," *Journal of Guidance, Control, and Dynamics*, Vol. 7, Nov.-Dec. 1984, pp. 695–702.

[9] Chalk, C.R., Neal, T.P., Harris, T.M., Pritchard, F.E., and Woodcock, R.J., "Background Information and Users Guide for MIL-F-8785B (ASG)," AFFDL-TR-69-72, Aug. 1969, pp. 417–461.

[10] Holley, W.E. and Bryson, A.E., "Wind Gust Modeling and Lateral Control for Automatic Landing," *Journal of Spacecraft and Rockets*, Vol. 14, Feb. 1977, pp. 65--72.

[11] Chen, C., *Introduction to Linear Systems Theory*, Hold, Rinehart, & Winston, New York, 1970.

Detection Filter Design: Spectral Theory and Algorithms

JOHN E. WHITE AND JASON L. SPEYER, FELLOW, IEEE

Abstract—**A new formulation of the detection filter problem is generated by assignment of the closed-loop eigenstructure under certain constraints. Detection filters, which are actually a specific class of observers, fix the output error direction of the system so that it can be associated with a particular failure mode and its known design failure direction. The derivation of detection filters from an eigensystem assignment approach permits a very transparent theory. The detection filter gains and closed-loop eigenvectors are obtained from a set of simultaneous equations. Necessary and sufficient conditions for the solution of these algebraic equations are determined which produce a complete theory for detection filters.**

I. Introduction

THE design of reliable, fault-tolerant control systems requires that system failures be detected and identified within acceptable time limits, such that the system feedback is not excessively corrupted. The principal tradeoff to be made in designing a redundancy management scheme is that of hardware redundancy versus the complexity and robustness problems of the software for analytic redundancy (i.e., combining the outputs of dissimilar devices through analytic kinematic and dynamic relationships to obtain redundancy). A survey on design methods for failure detection is given in [1]. Analytic redundancy management schemes are developed by forming and processing failure residuals. These residuals are essentially zero if no failure occurs and are nonzero if a failure occurs. The residual formation techniques in the literature may be categorized into two broad groups. Open-loop schemes [2] form one group. These schemes involve the construction of a set of parity equations which represent all of the analytical redundancies of a system. These parity equations are simply all of the possible input-output relationships of a given linear system. A generalized parity space [2] can be formed from the parity equations, and in the presense of a failure the resulting parity errors combine to provide a failure signature with directional characteristics in addition to the usual residual magnitude information. Theoretically, these directional signatures should facilitate the failure detection and identification process. However, the open-loop parity error characteristics are of a highly temporal nature and, therefore, the directional failure signature is not generally constrained. Furthermore, the failure magnitude of some or all of the parity residuals may disappear after n or fewer sample-times (n is the dimension of the state space). These problems would seem to limit the usefulness of the open-loop parity space concepts.

The second category of residual formation techniques is that of closed-loop schemes. Although any linear filter residual could be processed, one particular type of filter produces residuals with directional characteristics that can readily be associated with some known failure mode. These filters are known as detection filters, but are actually a particular class of observers. Unlike the directional failure signatures of the open-loop parity space method, detection filters act in a closed-loop fashion to fix the output direction associated with plant and actuator failures while restricting sensor failure output directions to lie in a plane. Furthermore, the output error magnitude never completely disappears when a failure has occurred.

The original theoretical development in detection filters was completed by 1973 [3], [4]. The intent of this paper is to reformulate the detection filter theory of [3], [4] as an eigensystem assignment problem. The algorithms of [3], [4] take the relatively indirect approach of generating a cyclic space. The current approach produces a straightforward derivation which yields a system of simultaneous linear equations to be solved for the detection filter gains and the closed-loop eigenvectors, once the closed-loop eigenvalues have been assigned. The detection filter terminology and certain referenced parts of this paper are taken from [3], [4]. Our results parallel those in multivariable control system design based on cnoosing the closed-loop eigenstructure to determine a unique feedback gain matrix [5]. Moore [5] has shown that, in addition to the usual freedom to choose the closed-loop eigenvalues, the closed-loop eigenvectors can be chosen from an m-dimensional subspace when there are m control inputs. Moore uses this flexibility in the choice of the eigenvectors to propose a design scheme for adjusting the distribution of the modes among the output components so as to shape the response characteristics of the system. This paper demonstrates that similar eigensystem assignment freedoms and design algorithms exist for a particular class of observers, known as detection filters, which can be completely defined by specification of a set of closed-loop eigenvalues along with appropriate constraints on the eigenvectors. Although there is a good deal of literature on eigenstructure assignment for both state and output feedback, the constraints imposed here on the observer gains require alternate derivations and algorithms. For additional detail, see [6].

In the next section definitions to establish notation and to introduce basic failure modeling considerations are presented. In Section III the algorithm for determining the detection gains and closed-loop eigenvectors is established. However, this algorithm assumes that the eigenvalues can be arbitrarily assigned. If a certain condition is not met, the algorithm must be further generalized. This is the topic of the subsequent sections. An example is used to illustrate all the theory. Although the detection filter was analyzed in [3] and [4], the proofs of the theorems and resulting algorithms for determining the detection gains and closed-loop eigenvalues by the eigensystem assignment method are generally different.

II. System Definition

The open-loop dynamic model in the absence of failures is given by

$$\dot{x} = Ax + Bu \tag{1}$$

where x is an $n \times 1$ state vector. The measurement equation in

Manuscript received July 21, 1986; revised February 12, 1987. This paper is based on a prior submission of October 5, 1984. Paper recommended by Associate Editor, H. L. Weinert.

The authors are with the Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, TX 78712.

IEEE Log Number 8714979.

0018-9286/87/0700-0593$01.00 © 1987 IEEE

the absence of sensor failures is written as

$$y = Cx \tag{2}$$

where y is an $m \times 1$ measurement vector. The detection filter is assumed to have the form of a linear filter such that

$$\dot{\hat{x}} = A\hat{x} + Bu + D(y - C\hat{x}) \tag{3}$$

where $\hat{x}$ is the state estimate and D is the detection gain. If the state error is defined as $\epsilon \triangleq x - \hat{x}$, then D is to be chosen such that the output error, $\bar{\epsilon} = y - C\hat{x}$, has restricted directional properties in the presence of a failure. The closed-loop dynamic equation becomes $\dot{\epsilon} = G\epsilon$ when there are no failures, where

$$G \triangleq A - DC. \tag{4}$$

The occurrence of a plant or actuator failure can usually be modeled by a single term added to (1) to produce

$$\dot{x} = Ax + Bu + f_i\mu_i \tag{5}$$

where f_i is the $n \times 1$ design failure direction associated with the ith plant or actuator failure, and μ_i is generally a time-varying scalar which may be a function of $x(t)$ or $u(t)$. *A priori* knowledge of μ_i is not required in the design of a detection filter and it is assumed that $\mu_i(t)$ is an arbitrary function of time. However, knowledge of the failure magnitude characteristics may be useful for distinguishing between different failures with the same output directions. For plant or actuator failures, the error system is rewritten as

$$\dot{\epsilon} = G\epsilon + f_i\mu_i, \qquad \bar{\epsilon} = C\epsilon. \tag{6}$$

The detection gain D will be determined in the following sections so that $\bar{\epsilon}$ is proportional to Cf_i in response to a failure corresponding to that modeled by the direction f_i. This output direction is maintained during the *transient* (assuming that the transients due to the initial conditions have settled-out before the failure occurs) and steady-state phases of the error response due to the system failure.

The occurrence of a sensor failure can generally be modeled with a single term added to (2) as

$$y = Cx + e_i\mu_i \tag{7}$$

where e_i is an $m \times 1$ unit vector corresponding to a failure in the ith sensor. For sensor failures the error system becomes

$$\dot{\epsilon} = G\epsilon - d_i\mu_i, \qquad \bar{\epsilon} = C\epsilon + e_i\mu_i \tag{8}$$

where d_i is the ith column of the detection filter gain matrix. The presence of d_i in (8) is a potential difficulty since the detection gain is not known *a priori*. The objective of the design procedure for a sensor failure is to determine two *a priori* directions associated with a failure in the ith sensor such that the output errors lie somewhere in the plane defined by Cd_i and e_i. Therefore, the closed-loop error system of (8) can be replaced by a system of the form

$$\dot{\epsilon} = G\epsilon - f_i^*\mu_i + f_i, \qquad \bar{\epsilon} = C\epsilon \tag{9}$$

where f_i is any direction such that $e_i \triangleq Cf_i$ and $f_i^* = Af_i$. It is shown in Section III-D that CAf_i lies in the plane generated by Cd_i and Cf_i.

The error system of (6) is used in the remaining analysis since it is easily generalized to sensor failures or other circumstances which are described by multiple failure directions.

III. Detection Filter Design

In this section the algorithm for determining the detection gain D is developed. First, the notion of detectability of a fault direction is defined. This definition requires that for arbitrarily placed filter poles, the gains be determined so that a failure direction induces a unique measurement residual direction. To ensure that the fault directions are detectable, certain general assumptions are imposed on the system. In later sections, some of these assumptions are removed. The restrictions imposed on the system by the assumptions and the requirement of detectability force groups of eigenvectors to produce output residual directions identical to those produced by the various fault directions. The number of eigenvectors associated with each output residual direction is determined by computing the dimension of the associated detection space. If the sum of the dimensions of all the detection spaces adds up to the dimension of the state space (this property is referred to as mutual detectability), then the filter eigenvalues can be arbitrarily assigned, and a simple algorithm for determining the detection gains and the closed-loop eigenvectors is developed. In Section VI techniques for making nonmutually detectable problems mutually detectable are developed.

A. Failure Detectability

The development of the theory of detection filters from the eigensystem assignment approach begins with a definition of the basic requirements for a detection filter. The definition of the detectability of a failure with the design direction f_i is given by Beard [3] as stated below.

Definition 1: The failure associated with f_i in the system described by (6) is detectable if there exists a filter gain matrix D such that

a) $\bar{\epsilon}(t)$ maintains a fixed direction in the output space, and

b) all eigenvalues of G can be arbitrarily specified, except for the constraint on the conjugate symmetry.

Condition a) forces the filter to have properties such that the output error direction $\bar{\epsilon}$ can be associated with the design error direction f_i. Condition b) is imposed so that the filter can be made stable, and also so that the response time of the filter can be adjusted. If condition b) is satisfied, the detection "filter" can also be used as a state estimation observer. The conjugate symmetry constraint will restrict the closed-loop eigenvectors associated with complex conjugate eigenvalues.

B. System Qualifications

The assumptions upon which the analysis of Sections III-C to III-E are based are

1) (A, C) is an observable pair.

2) $CF \triangleq [Cf_1, \cdots, Cf_r]$ is rank r.

3) $r = m$.

4) The closed-loop eigenvalues of G, λ_i, $i = 1, 2, \cdots, n$ are distinct.

The observability restriction is required for the usual state estimation reasons. The assumption that CF be rank r will be referred to as the condition of *output separability* [3]. The output separability condition produces a relatively uncomplicated closed-loop structure (Lemma 3). If $CA^jf_i = 0$ for $j = 0, 1, \cdots, \delta - 1$ and if $CA^\delta f_i \neq 0$, then all of the $Cf_i \neq 0$ assumptions and procedures can still be used if f_i is replaced everywhere by $A^\delta f_i$ [3], [4], [6]. Furthermore, if the output separability assumption is not satisfied because two failure directions of interest have identical output directions, the dynamics of the system may still allow for the detection of the two failure directions with a single detection filter. As a design procedure, one or both of the original directions can be replaced with A^jf_i for some $j > 0$ such that Assumption 2 is satisfied. The assumption is made in Sections III-C to III-E that $r = m$ since it is generally desirable to identify the maximum number of failures possible with a single detection filter. The $r < m$ case will be considered, however, in Section IV. The addition of a set of nonoutput-separable directions to F will be described in Section VII. Some constraints on eigenvalue assignability generally must be accepted to add these directions to the original set of r output separable directions. The analysis for the case where the eigenvalues are not all distinct will be given in Section V.

C. Detectability and Closed-Loop Eigenvector Constraints

The following analysis imposes the requirements of Definition 1 by placing constraints on the eigenstructure of the error systems. The distinct closed-loop eigenvalues λ_j and eigenvectors v_j of (4) are determined by

$$(\lambda_j I - G)v_j = 0 \tag{10}$$

for $j = 1, 2, \cdots, n$, where $Cv_j \neq 0$ since the system is assumed to be observable. The v_j are independent and span the error variable state space. The design failure vector can therefore be written as a linear combination of the closed-loop eigenvectors as

$$f_i = \sum_{j=1}^{n_i} \alpha_j^i v_j^i \tag{11}$$

where the i superscripts denote those closed-loop eigenvectors and coefficients which combine to form f_i, and $n_i \leq n$ is the number of nonzero α_j^i.

An important restriction on the eigenvectors v_j^i is given by the following theorem which is proved in Appendix A.

Theorem 1: Condition a) of Definition 1 is satisfied if and only if Cf_i and Cv_j^i are colinear for all values of $j = 1, 2, \cdots, n_i$.

Remark: It will be convenient to normalize the v_j^i such that $Cv_j^i = Cf_i$ for all $j = 1, \cdots, n_i$.

Since Theorem 1 can essentially be viewed as placing constraints on the closed-loop eigenvectors to achieve a unidirectional output error while condition b) of Definition 1 requires that the eigenvalues be arbitrarily assignable, the detection filter problem becomes that of solving the set of equations

$$\begin{bmatrix} \lambda_j^i I - A & D \\ C & 0 \end{bmatrix} \begin{bmatrix} v_j^i \\ w_i \end{bmatrix} = \begin{bmatrix} 0 \\ w_i \end{bmatrix} \tag{12}$$

for the detection gain D, where $w_i \triangleq Cf_i = Cv_j^i$ with $j = 1, 2, \cdots, \bar{n}_i \geq n_i$ and $i = 1, 2, \cdots m$. However, the number of eigenvalues and eigenvectors $\bar{n}_i$ to be assigned to the Cf_i constraints remains to be determined. The conditions under which a solution to (12) can be obtained for D and the v_j^i's will be given in Sections III-D and III-E. These conditions are the same as those required by Beard for the solution of his formulation of the problem. Interestingly, the appropriate results are easily derived here even though the formulation is different (i.e., (13) below differs from Beard's form of that equation).

D. Detection Spaces

The calculation of the detection gain and $\bar{n}_i$ with respect to a single design failure direction is now examined. Those equations of (12) which correspond to the nonzero α_j^i's in (11) can be summed to obtain

$$DCf_i = Af_i - \sum_{j=1}^{n_i} \alpha_j^i \lambda_j^i v_j^i \tag{13}$$

where the remark following Theorem 1 and (11) have been used to show that $\Sigma_{j=1}^{n_i} \alpha_j^i = 1$ for all $\alpha_j^i \neq 0$ corresponding to f_i. The form of the solution of (13) can be used to obtain certain information useful to the solution of (12).

Lemma 1: If D, S, and Q are matrices of dimension $n \times m$, $m \times r$, and $n \times r$, respectively, where $n \geq m \geq r$ and rank $(S) = r$, then the general solution of $DS = Q$, is given by

$$D = QS^+ + D[I - SS^+] \tag{14}$$

where D is an arbitrary $n \times m$ matrix and represents the freedom left in D after satisfying $DS = Q$, and $S^+ \triangleq (S^T S)^{-1} S^T$ is the Moore–Penrose pseudo left inverse [7].

A proof of this lemma is given in [3], [6]. The reader is referred to [8] for a more thorough discussion of the generalized inverse problem.

The solution of (13) is given by Lemma 1 as

$$D = \left(Af_i - \sum_{j=1}^{n_i} \alpha_j^i \lambda_j^i v_j^i\right)(Cf_i)^+ + D_i P(Cf_i) \tag{15}$$

where $P(S) \triangleq I - SS^+$ and $S = Cf_i$. Equation (15) cannot be employed to directly solve for D since the summation term is unknown. However, once D has been chosen so as to satisfy the detection filter constraint due to f_i in (13), a *new* system can be defined with a filter gain of D_i. This new system has the same form as the original problem and is useful because it allows the detection filter designer to determine how many eigenvalues can be arbitrarily specified by the choice of D_i, and the number of eigenvalues associated with Cf_i in (12). The new system is determined by substituting (15) into (4) to obtain

$$A - DC = A_i - D_i C_i \tag{16}$$

$$A_i \triangleq A - \left(Af_i - \sum_{j=1}^{n_i} \alpha_j^i \lambda_j^i v_j^i\right)(Cf_i)^+ C \tag{17}$$

$$C_i \triangleq P(Cf_i)C \tag{18}$$

and is characterized by the following lemma.

Lemma 2: If A_i, C_i, and D_i are real matrices of dimension $n \times n$, $m \times n$, and $n \times m$, respectively, the number of eigenvalues of $(A_i - D_i C_i)$ which can be arbitrarily specified by the free choice of D_i is equal to $q_i \triangleq$ rank (M_i) where

$$M_i \triangleq [(C_i)^T, (C_i A_i)^T, \cdots, (C_i A_i^{n-1})^T]^T. \tag{19}$$

The remaining $\nu_i \triangleq (n - q_i)$ eigenvalues of $(A_i - D_i C_i)$ are equal to the corresponding eigenvalues of A_i, which also are those eigenvalues of G associated with f_i. The proof is given in Appendix B.

Remark: Observe that Lemma 2 is *not* written in an implementable form for this problem because (17) contains the v_j^i's which are unknown. However, A_i in (19) can be replaced by $K_i \triangleq A[I - f_i(Cf_i)^+ C]$. The equivalence between A_i and K_i can be established by using $C_i K_i^k v_j^i = 0$ for $k = 0, 1, \cdots, n - 1$ [6].

Definition 2: The null space of M_i is the detection space of f_i.

Definition 3: The dimension of the detection space of f_i is defined to be the *detection order* ν_i of f_i, where $\nu_i = n - q_i$.

Definition 4: The failure vector f_j is *detection equivalent* to f_i if

a) every detection filter for f_i is also a detection filter for f_j, and

b) $Cf_i = \beta Cf_j$ (Assumption 2 of Section III-B implies that $Cf_i \neq 0$ and $Cf_j \neq 0$) where β is any nonzero constant.

The detection space is a G-invariant subspace of the error variable state space which represents that part of the system affected by f_i or some detection equivalent direction. The invariance property is clear from (13) since the summation term represents some vector in the detection space of f_i and the other two terms can be combined to form Gf_i. This invariance property implies that the controllable space of f_i with respect to G, W_i given in (A-1), is a subspace of the detection space of f_i, since W_i is the smallest G-invariant subspace containing f_i. The fact that W_i is generally a proper subspace of the detection space is the result of the maximum rank of W_i being constrained to be $n_i \leq \nu_i$. This constraint on the rank of W_i can be observed from the substitution of (11) and (10) into W_i.

The detection space of f_i contains f_i and all of the f_j which are detection equivalent to f_i, since $C_i f_i = 0$ and $K_i f_i = 0$ imply that f_i and the detection equivalent f_j's lie in the null space of M_i.

Lemma 2 and Definition 3 imply that there are ν_i eigenvalues whose corresponding eigenvectors span the detection space of f_i. Since n_i eigenvectors of (11) are known by Theorem 1 to lie in the detection space, then $\nu_i \geq n_i \geq 1$. Therefore, since there are ν_i eigenvectors which satisfy the collinear constraint of Theorem 1 for all detection equivalent f_j's, then $\bar{n}_i = \nu_i$ is the number of eigenvalues and eigenvectors to be assigned to the Cf_i constraint in solving the algebraic equation (12).

It remains to show that all n eigenvalues can be arbitrarily assigned as required by Definition 1.

Definition 5: The vectors in the set F are mutually detectable if there exists a D which satisfies the detectability conditions of Definition 1 for all f_i in the set F.

The condition for the set F to be mutually detectable is given in Theorem 2 for the case when $r = m$.

Theorem 2: If the set of vectors in F are output separable, then the f_i's in F are mutually detectable if and only if

$$\sum_{i=1}^{m} \nu_i = n.$$

The proof is given in Section IV where the $r < m$ case is discussed. Procedures for making a system mutually detectable when the condition in Theorem 2 fails is given in Section V.

E. Solution of the Algebraic Equation (12)

The following theorem states the conditions under which (12) can be solved for the detection gain D and the closed-loop eigenvectors. The proof is given because it is constructive in developing an algorithm for the solution to (12).

Theorem 3: Given the system qualifications of Section III-B, then the condition $\Sigma_{i=1}^{m} \nu_i = n$ implies that the system of (12) can be solved for the detection gain matrix D and the closed-loop eigenvectors, v_j^i, where $j = 1, 2, \cdots, \nu_i$.

Proof: By Theorem 2 the n eigenvalues required in (12) can be arbitrarily specified ν_i at a time, while simultaneously being associated with a particular Cf_i constraint. Therefore, for each f_i, n linear combinations of the elements of D can be determined along with the $n\nu_i$ elements of v_j^i's. To determine these unknowns, there are $n\nu_i$ eigenvector equations and $m\nu_i$ eigenvector constraints. When $m\nu_i \geq n$, then those equations in (12) corresponding to f_i can be used to completely solve for the corresponding v_j^i's and the associated set of n constraints on D. This is possible because there are an equal number of *independent* equations and unknowns. However, if $m\nu_i < n$, then v_j^i is representable as a linear combination of any basis for the ith detection space. Since the ith detection space must be orthogonal to M_i of (19), a basis set for the detection space can be found by computing the unobservable subspace of the (C_i, K_i) system [9]. Hence, v_j^i can be written as $v_j^i = \Omega_i\beta_j^i$ where Ω_i is an $n \times \nu_i$ matrix whose columns form a basis for the detection space with respect to f_i, and β_j^i is a $\nu_i \times 1$ vector of coefficients. It is now shown that enough independent equations will exist for solution of the elements of the $v_j^i = \Omega_i\beta_j^i$ and the n constraints on D. The number of equations in (12) remains at $n\nu_i + m\nu_i$ while the number of unknowns has been reduced to $\nu_i^2 + n$. If the number of equations can be shown to be greater than or equal to the number of unknowns, it will be possible to solve for the β_j^i and the constraints on D. It will now be assumed that this is true so that $[(n + m)\nu_i] \geq [\nu_i^2 + n]$ which will be rewritten so that the validity of the inequality is clear. When $r \leq m$, then $n \geq \nu_i$. This implies that $n = \nu_i + c$ where $c \geq 0$. Hence, $[\nu_i^2 + (m + c)\nu_i] \geq [\nu_i^2 + \nu_i + c]$, and then $[m + c]\nu_i \geq [\nu_i + c]$. Since $m \geq 1$ and $\nu_i \geq 1$, then the inequality is valid. Hence, solution of (12) is possible and Theorem 3 is proven.

Remark: As discussed in Section II, sensor failures are included by determining *a priori* fault directions f_i and f_i^* such that $Cf_i = e_i$, and Cf_i^* lie in the plane composed of Cf_i and Cd_i. Note that for $f_i^* \triangleq Af_i - \eta\hat{f}_i$ where $\hat{f}_i$ lies in the detection space of f_i and η is an arbitrary constant, Cf_i^* lies in the plane composed of $CDCf_i = Cd_i = CAf_i - \Sigma_{j=1}^{\nu_i} \alpha_j^i\lambda_j^i e_i$ and e_i where (13) is used.

IV. $r < m$ FAILURE DIRECTIONS

The assumptions of Section III-B are relaxed to allow $r < m$. The condition for mutual detectability of Theorem 2 must be generalized. This is done by determining the "detection space" associated with the set $F \triangleq \{f_1, \cdots, f_r\}$. In a manner analogous to that for the single failure, the detection gain for multiple failures must simultaneously satisfy r equations of the form of (13). This set of equations can be written in matrix form as $DCF = Q_d$ where the columns of Q_d are the right-hand sides of (13) for $i = 1, 2, \cdots, r$. Lemma 1 gives the solution of $DCF = Q_d$ as

$$D = Q_d(CF)^+ + DP(CF) \tag{20}$$

where

$$A - DC = A - DC, A \triangleq A - Q_d(CF)^+C \tag{21}$$

$$C \triangleq P(CF)C, K \triangleq A[I - F(CF)^+C]. \tag{22}$$

As per Lemma 2 and the following remark, the observability matrix with respect to the system (A, C) is

$$M \triangleq [(C)^T, (CA)^T, \cdots, (CA^{n-1})^T]^T. \tag{23}$$

The number of eigenvalues which are freely assignable by D is given by Lemma 2 to be rank $(M) \triangleq q$. The number of eigenvalues associated with making D a detection gain for the set of r failure vectors in F is therefore given by $\nu = n - q$.

Definition 6: The dimension of the null space of M is $\nu = n - q$ and is defined to be the *group detection order* of the set $F = \{f_1, \cdots, f_r\}$.

For F defined by a set of $r \leq m$ failure directions Theorem 2 generalizes to

Theorem 4: The f_i's in F are mutually detectable if and only if

$$\nu = \sum_{i=1}^{r} \nu_i. \tag{24}$$

Proof is given in Appendix C. Note that if $m = r$, then CF is invertible and $C = 0$ in (22). Therefore, $\nu = n$, and Theorem 4 implies Theorem 2.

The results of Theorem 3 must be slightly modified since q eigenvalues and eigenvectors remain to be assigned after the detection filter has been designed for the f_i, $i = 1, 2, \cdots, r$. q is the rank of M, which is defined by (23). These q eigenvalues and eigenvectors are freely assignable, provided that the eigenvectors are independent of those which span the r detection spaces. All of the solution techniques of Theorem 3 are also applicable for the case of $r \leq m$, since the developed techniques allow for the independent solution of the subset of equations in (12) corresponding to a single f_i for the appropriate ν_i eigenvectors and n constraints on D [6].

V. DETECTION FILTERS WITH NONDISTINCT EIGENVALUES

The assumptions of Section III-B are also imposed on the analysis of this section, except that nondistinct eigenvalues and $r \leq m$ are now allowed.

Some interesting eigenstructure constraints are imposed by output separability when nondistinct eigenvalues are allowed. First, it is demonstrated that the detection spaces are independent.

Lemma 3: If the failure vectors in F are output separable, then the detection spaces of the f_i's are pairwise independent. The proof is given in Appendix D.

The proof shows that there are no eigenvectors which span some overlap in the detection spaces. The implication of this lemma on the Jordan form structure of G is that any Jordan block must be completely contained in the ν_i-dimensional part of the Jordan structure corresponding to the ith detection space, even if the algebraic multiplicity of the eigenvalue is greater than ν_i. This requires that whenever identical eigenvalues are to be assigned to k detection spaces, the eigenvalue must be assigned a geometric multiplicity of k to preserve the independence of the detection spaces. This can be illustrated quite simply by falsely assuming that v_1 and v_2 lie in different detection spaces when v_2 is a generalized eigenvector of v_1. By the definition of a generalized eigenvector, then it must be true that $CGv_2 - \lambda Cv_2 = Cv_1$ which must be equal to (α is a constant scalar coefficient) $\alpha Cf_2 = Cf_1$ if v_2 and v_1 are to lie in their respective detection spaces. However, this is a contradiction by the definition of output separability.

The fact that the geometric multiplicity must be exactly k can be shown by assuming falsely that two primary eigenvectors within a single detection space have the same eigenvalue. If these eigenvectors are denoted by v_1 and v_2, then $v_1 - v_2$ must also be an eigenvector. However, this implies a contradiction of the observability assumption since $C(v_1 - v_2) = 0$ by Theorem 1. Hence, there can only be one primary eigenvector per detection space associated with a particular eigenvalue.

If the system of (12) is rewritten to include generalized eigenvectors, then the results of Sections III-C to III-E can be confirmed to be valid when nondistinct eigenvalues are allowed since Theorem 1 can be proven [6] for an f_i defined in terms of primary and generalized eigenvectors. The results of Sections III-D and III-E can also be extended to the nondistinct eigenvalue case by inspection if the summation term in (13) is modified. Because of Theorem 1, this modified term will drop out of an analysis similar to that used to obtain M_i in terms of K_i rather than A_i (see remark after Lemma 2), and the results of those sections carries over to the nondistinct eigenvalue problem.

The system of (12) is still valid for those values of j associated with primary eigenvectors, while the right side of (12) must be replaced by $[-(v'_{j-1})^T, w_i^T]^T$ for those values of j corresponding to generalized eigenvectors. The results of Section IV are trivially extended to the nondistinct eigenvalue case.

VI. Nonmutually Detectable Failures

If the condition of (24) is not satisfied, then a solution to the detection filter problem cannot be found which is detectable in the sense of Definition 1. In this case, $(\Sigma_{i=1}^r \nu_i) + q < n$ and there exists an excess subspace of dimension ν_e such that

$$\nu_e \triangleq n - \left[\left(\sum_{i=1}^{r} \nu_i\right) + q\right] = \nu - \sum_{i=1}^{r} \nu_i. \tag{25}$$

This excess subspace exists as a result of making the detection filter respond to the set of failure directions in F, which requires ν eigenvalues, while only being able to freely assign $\Sigma_{i=1}^r \nu_i$ eigenvalues with respect to the individual failure directions. Hence, ν_e eigenvalues will be fixed by the choice of the system and the set F. When this occurs the system is said to be nonmutually detectable [3] or restrictive [10].

The procedure [6] for obtaining a mutually detectable system by eliminating some of the f_i's in F essentially examines the eigenvectors which span the excess subspace so as to associate the removal of an f_i with the elimination of some of those excess eigenvectors. After examining each of the f_i in such a manner, the dimension of the excess subspace for all possible combinations of the f_i can be simply determined. Those combinations which produce an excess subspace of dimension zero are those subsets of the original set F which permit a mutually detectable problem. Note that (24) must be satisfied when $r = 1$ because $M_i = M$ and a mutually detectable system always occurs. The removal of failure directions from the original set of r directions may be averted by an alternate technique which was first suggested by Jones [4]. This technique is that of adding dynamics to the original open-loop system in such a way that (24) is satisfied. If the original open-loop system is a minimal realization, then it is both observable and controllable (with respect to B). Although the new open-loop system will be required to be an equivalent realization of the original system, the enlarged system will be observable but not controllable.

The motivation for enlarging the state space to obtain a mutually detectable set of failures requires a slightly more detailed understanding of the structure of the excess subspace. Cv_e, where v_e is any one of the eigenvectors spanning the excess subspace, does not lie in any one detection space. However, since $Cv_e \neq 0$ by the discussion with respect to (10) and is directly observable, then Cv_e must then be some linear combination of the directions of Cf_i for $i = 1, 2, \cdots, r$. Each eigenvector in the excess subspace must have a component in two or more of the detection spaces, otherwise by Definition 4 and Theorem 1 the eigenvector would lie in and span some detection space. The removal of a failure f_k from the set F removes those eigenvectors of the excess subspace which have components along Cf_k. The removal of some other failure direction other than f_k may also remove some of the same eigenvectors that f_k would remove. The essential idea of state-space enlargement is to increase the dimension of the state space in such a way that the presence or absence of an f_k in F does not affect the dimension of the excess subspace of the enlarged state space. This will be accomplished by the choice of a new open-loop system matrix such that the kth detection space is enlarged by an amount equal to the number of eigenvectors of the original excess subspace which have components along Cf_k, while ensuring that the new excess subspace eigenvectors have no component along Cf_k.

Before other constraints on the enlargement of the state space can be considered, the requirements for an observable, equivalent realization $(\tilde{A}, \tilde{B}, \tilde{C})$ with a dimension of $\tilde{n} > n$ must be established and satisfied. The open-loop models of (A, B, C) and $(\tilde{A}, \tilde{B}, \tilde{C})$ are defined to be equivalent realizations (i.e., input-output equivalent) if $CA^jB = \tilde{C}\tilde{A}^j\tilde{B}$ for $j \leq \tilde{n}$ when $\tilde{n} > n$. One form of the system $(\tilde{A}, \tilde{B}, \tilde{C})$ which can be observable and is input-output equivalent for any $\tilde{n} > n$ is given by

$$\tilde{C} = [C \quad 0], \ \tilde{A} = \begin{bmatrix} A & A_{12} \\ 0 & A_{22} \end{bmatrix}, \ \tilde{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}. \tag{26}$$

Let ν_{ek} and ν_{eR} represent the number of eigenvectors of the original excess space which have output components that lie and do not lie, respectively, along the direction Cf_k. The sum of these quantities is ν_e. The state space enlargement approach is made possible by the following theorem which is proved in Appendix E.

Theorem 5: There exists an observable extension of (A, C) into $(\tilde{A}, \tilde{C})$ of the form of (26) for $\tilde{n} \triangleq n + \nu_{ek}$ such that a) $\tilde{\nu}_{ek} = 0$, b) $\tilde{\nu}_k = \nu_k + \nu_{ek}$, c) $\tilde{\nu}_j \geq \nu_j$ for all $j \neq k$, and d) $\tilde{\nu}_e \leq \nu_e$. Sufficient conditions for a) through d) to occur are that A_{12} and A_{22} be chosen as

$$A_{12} \triangleq [-\sigma_k^1(\lambda_{ek}^1 f_k - \hat{f}_k), \cdots, -\sigma_k^N(\lambda_{ek}^N f_k - \hat{f}_k)], \qquad A_{22} \triangleq \Lambda_{ek} \tag{27}$$

where $N \triangleq \nu_{ek}$, $\hat{f}_k \triangleq$ summation term of (13) or the analogous nondistinct eigenvalue summation term, and λ_{ek}^j, Λ_{ek} are the eigenvalues in element and matrix (either diagonal or Jordan form) representations that correspond to the eigenvector v_{ek}^j and eigenmatrix V_{ek}, respectively.

Remark: The scalar σ_k^j's of (27) may be calculated as $\sigma_k^j = k$th row of $(CF)^+Cv_{ek}^j$.

A sequential application of the Theorem 5 can be used to generate a mutually detectable system. Repeated application of those results with respect to successive failure directions in F will cause each of those directions in turn to be eliminated from those eigenvectors which remain to span the excess subspace. Eventually, the excess subspace will be eliminated entirely and a mutually detectable system will have been formed.

Theorem 6: The nonmutually detectable system of (A, C, F) with output separable f_i's can be enlarged into an observable, equivalent realization $(\tilde{A}_r, \tilde{C}_r, \tilde{F} \triangleq [F^T, 0]^T)$ which is mutually detectable with respect to the same set of failure directions (the $\tilde{f}_k$'s) with the minimal state-space dimension of

$$\tilde{n}_r = n + \sum_{k=1}^{r} \nu_{ek} - \nu_e \tag{28}$$

after $r - 1$ applications of Theorem 5.

The proof is given in Appendix F.

VII. Output Stationarity

If the freedom to arbitrarily choose the closed-loop eigenvalues is somewhat restricted, [4] has shown that more than m failure directions can be designed into the detection filter. By definition, these additional directions are not output separable with the original set of directions in F. If h_k is a direction to be added to the set of m directions in F, then

$$h_k = \sum_{i=1}^{m} \delta_i f_i + \xi_k \tag{29}$$

where some of the δ_i may be zero and $C\xi_k \triangleq 0$. If h_k satisfies certain conditions, then the output direction associated with h_k can be made unidirectional by requiring that particular subsets of the closed-loop eigenvalues take on identical values. In the terminology of [4] this is referred to as making h_k output stationary with F. The simplest case of this occurs when $m = n$ and all of the eigenvalues are chosen to be identical. Under these conditions, $\nu_i = 1$ for $i = 1, 2, \cdots, n$. Since $\nu_i = 1$, then $v_i = f_i$ and the eigen equations become $Gf_i = \lambda f_i$. Here h_k is just a linear combination of the f_i's since $\xi_k \triangleq 0$ is required for $C\xi_k = 0$ to be true. Now any h_k can be detected by the detection filter for F since every h_k will also be an eigenvector for G.

A. Output Stationarity for a Single Additional Failure Direction

The fully measurable case is a powerful motivation for examining the general conditions under which additional failure directions can be detected. This section will determine those conditions which must be satisfied to make a single additional failure direction h_k output stationary with the failure directions of the set F.

The following assumptions are made for the analysis to follow: 1) $h_k \neq 0$, 2) $r = m$, 3) A, C, F imply mutual detectability, and 4) distinct eigenvalues. The first assumption follows directly from the previous assumption that $Cf_i \neq 0$, (29), and output separability. The second assumption is made so that the maximum number of output separable failures will be designed into the detection filter. This in turn assists in maximizing the number of h_k's which can be made output stationary with the directions in F. Extension of the results to the $r < m$ case will follow trivially from the $r = m$ analysis. The third assumption requires that the system be mutually detectable, either naturally or by the methods of Section VI. This assumption implies that no unassignable eigenvalues exist as a result of fixing the output directions in CF. This allows for the maximum possible flexibility in determining the conditions for output stationarity. Similarly, distinct eigenvalues are assumed because this aids in maximizing the freedom allowed in specifying output stationarity criteria. Complete eigenvalue assignability is the freedom sacrificed in fixing more than m failure directions.

The output stationarity problem is concerned with fixing the h_k detection subspace in the state space while simultaneously maintaining the detection subspaces with respect to the directions in F. The detection space for some direction h_k, which is to be made output stationary with the output separable directions in F, is defined to be of some dimension $\bar{\nu}_k$, where $\bar{\nu}_k \leq \min(\nu_i)$ for all i's corresponding to nonzero δ_i's in (29) [6]. The requirements for output stationarity are summarized in Theorem 7. The proof is given in Appendix G.

Theorem 7: For a mutually detectable problem defined by A, C, F, it is possible to make the nonoutput-separable direction h_k output stationary with the directions in F if and only if two conditions are satisfied. First, ξ_k of (29) must lie in the union of the detection spaces of the f_i's which correspond to nonzero δ_i (assume for notational simplicity that these correspond to $i = 1, 2, \cdots, l$). The second condition is that each group of ν_i eigenvalues must contain the same set of $\bar{\nu}_k$ arbitrarily assignable eigenvalues for $i = 1, 2, \cdots, l$. If $\bar{\nu}_k < \nu_i$ for $i = 1, 2, \cdots, l$, then $\nu_i - \bar{\nu}_k$ unassignable eigenvalues will exist with respect to the ith detection space.

Remark: Implementation of the results of Theorem 7, in the spirit of the previous algorithms, is quite straightforward. The first case to be considered is when $\bar{\nu}_k = \nu_i$ for all $i = 1, 2, \cdots, l$. In this case the determination of the detection gain matrix and the closed-loop eigenvectors proceeds exactly the same way as with (12), except that the eigenvalues with respect to the detection spaces of the f_i's for $i = 1, 2, \cdots, l$ must be chosen as identical sets of $\bar{\nu}_k$ eigenvalues. The eigenvalues with respect to the other detection spaces can be freely assigned.

The second case to be evaluated is when $\bar{\nu}_k < \nu_i$ for some or all of the f_i detection spaces for $i = 1, 2, \cdots, l$. In this case the equations of (12) must be solved with the same eigenvalue constraints as in the first case. However, $\nu_i - \bar{\nu}_k$ eigenvalues are unassignable for each of the detection spaces where $\nu_i > \bar{\nu}_k$. These unassignable eigenvalues complicate the solution of (12) since there will be more unknowns than equations. This complication is eliminated by requiring that the system equations of (12) be solved simultaneously with

$$h_k = \sum_{i=1}^{l} \sum_{j=1}^{\nu_i} \delta_j^i v_j^i, \quad \delta_j^i \triangleq \delta_i \alpha_j^i + \gamma_j^i. \tag{30}$$

$$Gh_k = \sum_{i=1}^{l} \sum_{j=1}^{\nu_i} \delta_j^i \lambda_j^i v_j^i, \quad \sum_{j=1}^{\nu_i} \delta_j^i = \delta_i \tag{31}$$

along with the condition that $\delta_j^i = 0$ for all $j > \bar{\nu}_k$ and $i = 1, 2, \cdots, l$. The δ_i's of (30) and (31) are known *a priori* from (29) and γ_j^i is found from the representation $\xi_k = \Sigma_{i=1}^{l} \Sigma_{j=1}^{\nu_i} \gamma_j^i v_j^i$. Equation (31) requires that h_k satisfy an equation of the form of (13) with the same detection gain matrix as that used to fix the output directions in F. Also, all of these equations are expressed in terms of the eigenvalues and eigenvectors of the f_i detection spaces and are compatible with the unknowns of (12).

VIII. Examples of Detection Filter Design

Example (a): This is an example of the eigensystem assignment methods for detection filter design when the closed-loop eigenvalues are chosen to be distinct, $r = m$, and the A, C, F system is mutually detectable. Let

$$A = \begin{bmatrix} 0 & 3 & 4 \\ 1 & 2 & 3 \\ 0 & 2 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad f_1 = \begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix}, \quad f_2 = \begin{bmatrix} 1 \\ -1/2 \\ 1/2 \end{bmatrix}.$$

Since the rank of CF is $r = 2$, then the failure directions of f_1 and f_2 are output separable. The test for mutual detectability using Lemma 1 and the following remark produces M_i for $i = 1, 2$ as

$$M_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 5 \\ 0 & 0 & 25 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 7/2 & 5/2 \\ 7/2 & 91/4 & 63/4 \end{bmatrix}$$

where the zero or redundant rows have been omitted. The rank of M_1 is 1 which implies that $\nu_1 = n - 1 = 3 - 1 = 2$. The rank of M_2 is 2 and this means that $\nu_2 = n - 2 = 1$. Since $n = \nu_1 + \nu_2$, then the system A, C, F is mutually detectable. The closed-loop eigenvalues are now assigned to the respective detection spaces. Since $\nu_1 = 2$, the choices of $\lambda_1^1 = -2$ and $\lambda_2^1 = -3$ are assigned to the detection space of f_1. Likewise, the choice of $\lambda_1^2 = -4$ is made for the detection space of f_2. The closed-loop eigenvectors and the detection gain, D, are now determined from (12). The two sets of linear equations obtained for λ_1^1 and λ_2^1 can be solved to obtain $d_{11} = 9$, $d_{21} = 7$, $d_{31} = 2$, while the two eigenvectors which span the detection space of f_1 are calculated to be $v_1^1 = [3\ 1\ 0]^T$, $v_2^1 = [2\ 1\ 0]^T$. The single equation of (12) for λ_1^2 produces the detection gain elements as $d_{12} = 18$, $d_{22} = 6$, and $d_{32} = 9$ where the eigenvector is known to be f_2 since $\nu_2 = 1$.

Example (b): This example demonstrates the techniques of Section V for $r = m$, nondistinct eigenvalues, and a mutually detectable system. This problem is a repetition of Example (a) except that here all of the eigenvalues are chosen to be identical. Because of the detection space structure of this example, λ will have a geometric multiplicity of 2 but an algebraic multiplicity of 3, where $\lambda = -2$. The system of (12) is used for λ_1^1, while for λ_2^1 (12) must be modified to accommodate generalized eigenvectors as discussed in Section V. These two sets of linear equations can be solved to produce $d_{11} = 7$, $d_{21} = 6$, $d_{31} = 2$, and $v_1^1 = [3\ 1\ 0]^T$. For λ_1^2 (12) can be used where again the eigenvector is constrained to be f_2 since $\nu_2 = 1$. The detection gain elements are $d_{12} = 12$, $d_{22} = 7$, and $d_{32} = 7$.

Example (c): This example deals with the implementation of the detection algorithms of Section VI for the restrictive problem where distinct eigenvalues and $r = m$ are assumed. The system of Example (a) is used again here, except that $f_1 = [0\ 0\ 1]^T$. From Lemma 1 and the following remark $\nu_1 = 1$. This problem is, therefore, restrictive since $\nu > \nu_1 + \nu_2 = 2$. The dimension of the excess subspace is $n = 3 - 2 = 1$. Since only two eigenvalues are arbitrarily assignable, $\lambda_1^1 = -2$ and $\lambda_1^2 = -3$ with the excess eigenvalue left to be determined. The two systems of equations from (12) are solved for the detection gain matrix where the first column of D is $[-3\ 3\ 1]^T$ and the second column of D is $[4\ 3\ 7]^T$. From G of (4) calculate v_e and λ_e from (10) as $\lambda_e = 2$ and $v_e = [12\ 4\ 1]^T$.

The results of Theorem 5 may be employed to enlarge the state space where $\tilde{f}_1 = \lambda_1 v_1^1 = -2[0\ 0\ 1]^T$. The enlargement technique is applied with respect to f_1 in this example, although f_2 could just have easily been chosen. Now A_{12} of Theorem 5 can be written as $A_{12} = -5[2f_1 + 2f_1] = [0\ 0\ -20]^T$ while $A_{22} = \lambda_e = +2$. This choice of the new open-loop system must enlarge the dimension of the detection space of f_1 by b) of Theorem 5, and must also enlarge the detection space of f_2 by Theorem 6. This problem is an example of the case where $\tilde{\nu}_e < \nu_e$. The dimension of both detection spaces will be enlarged by one and, hence, the new system will be mutually detectable.

Example (d): This example will demonstrate the techniques of Section VII. The objective here will be to make a single additional nonseparable direction $h_1 = [0\ 0\ 1]^T$ output stationary with the failure directions of Example (a). By definition the direction h_1 can be written in the form of (29) as $h_1 = f_1 + 2f_2 + \xi_1$ where $\delta_1 \triangleq 1$, $\delta_2 \triangleq 2$, and $\xi_1 \triangleq [1\ 0\ 0]^T$. Because the h_1 direction coincides with the f_1 direction of Example (c), the calculation of $\bar{\nu}_1$ here is unnecessary since the calculation of ν_1 in Example (c) implies that $\bar{\nu}_1 = 1$. From Example (a) it should be recalled that $\nu_1 = 2$ and $\nu_2 = 1$. Hence, the application of Theorem 7 will require that only one freely assignable eigenvalue be assigned to the detection spaces of f_1 and f_2, and it must be identical for both detection spaces. Because $\nu_1 > \bar{\nu}_1$ there will be one unassignable eigenvalue with respect to the detection space of f_1. The other condition of Theorem 7 which must be satisfied is that ξ_1 lie in the union of the detection spaces of f_1 and f_2. The requirement is clearly satisfied since the eigenvectors which span the detection spaces of f_1 and f_2 will also span the state space and, therefore, ξ_1 can be written in terms of those eigenvectors. As mentioned in Section VII-A, the calculation of the eigenvectors and detection gain is complicated by the additional equations to be satisfied when making some vector output stationary with the directions in F. The system of (12) can be employed to write equations with respect to $\lambda_1^1 = -2$, $\lambda_2^1 =$ unassignable, and $\lambda_1^2 = -2$. The additional equations to be utilized for the output stationarity case are (30) and (31). The additional constraint that $\bar{\delta}_2^1 = 0$ must also be enforced since $\nu_1 > \bar{\nu}_1$. All of these equations can be solved to obtain

$$D = \begin{bmatrix} -1 & 4 \\ 2 & 3 \\ 2 & 7 \end{bmatrix}, \quad V = \begin{bmatrix} -2 & 2 & 1 \\ 1 & 1 & -1/2 \\ 0 & 0 & 1/2 \end{bmatrix} \tag{32}$$

with $\bar{\delta}_1^1 = 1$, $\bar{\delta}_2^1 = 0$, $\delta_1^2 = 2$, and $\lambda_2^1 = +2$.

The reader should notice that the unassignable eigenvalue here takes on the same value as the excess eigenvalue of Example (c) where h_1 and f_2 of the above example were used as the design failure directions. This is an interesting check of the above output stationarity procedure, since any effort to design the directions of h_1 and f_2 into a detection filter must result in an unassignable eigenvalue with a value of +2. The comparison between this example and that of Example (c) may be further enhanced if Example (c) is recalculated with $\lambda_1^2 = -2$. In this case, the gain matrix and excess eigenvalue are identical to the gain matrix and unassignable eigenvalue of Example (d). Furthermore, v_e of the recalculated version of Example (c) will be identical to v_2^1 of Example (d), which is associated with the unassignable eigenvalue λ_2^1. Therefore, the state enlargement technique could be used as in Example (d) to produce a mutually detectable system.

VIII. Conclusions

A derivation of the detection filter theory from an eigensystem assignment approach has been presented. The motivations for and the development of this theory are easily accomplished by this approach. The analysis results in a set of simultaneous equations to be solved for the detection filter gains and the closed-loop eigenvectors. Necessary and sufficient conditions for the solution of this system of equations have been given. An example is presented which illustrates and integrates all of the theory and associated algorithms.

Appendix A

Proof of Theorem 1

Theorem 1 has a two-part proof which is given below in the form of two lemmas and their respective proofs.

Lemma: Condition a) of Definition 1 is satisfied if and only if rank $(CW_i) = 1$.

Proof: This proof is due to Beard [3]. The controllable space of f_i with respect to G is spanned by the columns of the controllability matrix W_i, where W_i is defined as

$$W_i \triangleq [f_i, Gf_i, \cdots, G^{n-1}f_i]. \tag{A.1}$$

Sufficiency of the lemma can be established by noting that $\epsilon(t)$ must lie in the range space of W_i. Therefore, $\epsilon(t)$ can be written as a linear combination of the columns of W_i as $\epsilon(t) = W_i g(t)$, where $g(t)$ is an $n \times 1$ vector of coefficients. The output error can now be written as $\bar{\epsilon}(t) = C\epsilon(t) = CW_i g(t)$. It is sufficient that CW_i be of rank 1 to ensure that $\bar{\epsilon}(t)$ is unidirectional for any $g(t)$. The necessity of the lemma can be shown by observing that $\epsilon(t)$ can be driven by $\mu_i(t)$ to any state in the controllable space of f_i with respect to G. Hence, condition a) of Definition 1 is guaranteed for arbitrary $\mu_i(t)$ only if rank $CW_i = 1$. □

Lemma: Rank $(CW_i) = 1$ if and only if Cf_i and Cv_j^i are collinear for all $j = 1, 2, \cdots, n_i$.

Proof: The i superscripts on the v_j^i's and the corresponding λ_j^i's and α_j^i's have been suppressed in this proof for the sake of notational simplicity. Sufficiency can be demonstrated by replacing the f_i in CW_i' with (11), and then (10) can be employed to obtain

$$CW_i'=\left[\sum_j \alpha_j Cv_j,\ \sum_j \alpha_j \lambda_j Cv_j,\ \cdots,\ \sum_j \alpha_j \lambda_j^{n-1} Cv_j\right]. \quad \text{(A.2)}$$

Now if Cv_j and Cf_i are collinear for all j such that $\alpha_j \neq 0$, then CW_i' becomes

$$CW_i=[\boldsymbol{\alpha}_1 Cf_i,\ \boldsymbol{\alpha}_2 Cf_i,\ \cdots,\ \boldsymbol{\alpha}_{n-1} Cf_i] \quad \text{(A.3)}$$

where the $\boldsymbol{\alpha}_k$ represent the coefficients of Cf_i after the summations have been executed. Clearly, the rank of CW_i is one and sufficiency has been proved.

Necessity can be proven if it can be shown that $\bar{\epsilon}(t)$ is unidirectional only when the conditions of the lemma are satisfied. The solution to (6) can be written as

$$\epsilon(t)=e^{Gt}\epsilon(0)+\int_0^t e^{G(t-\tau)} f_i \mu_i(\tau)\, d\tau. \quad \text{(A.4)}$$

The transient in (A.4) is considered to be zero, by assuming that either $\epsilon(0) = 0$ or that the transient due to the initial condition has settled out. (This requires that G be chosen to be stable.) Then, the substitution of (11) into (A.4) and the assumption that μ_i is constant (any conditions obtained under this assumption must also apply for an arbitrary $\mu_i(t)$ since μ_i = constant is still a possible failure mode) gives

$$\epsilon(t)\approx\sum_{j=1}^{n_i} \mu_i \alpha_j \int_0^t e^{G(t-\tau)}\, d\tau\, v_j. \quad \text{(A.5)}$$

If (A.5) is premultiplied by C, $\exp[G(t - \tau)]v_j = \exp[\lambda_j(t - \tau)]v_j$ is substituted into (A.5), and the integration of (A.5) is performed, then the output error becomes

$$\bar{\epsilon}(t)\approx -\mu_i \sum_{j=1}^{n_i}\left(\frac{\alpha_j}{\lambda_j}\right)[1-e^{\lambda_j t}]Cv_j. \quad \text{(A.6)}$$

By (A.6) the conditions of the lemma are necessary for $\bar{\epsilon}(t)$ to have a fixed direction and, therefore, for CW_i to be rank 1. □

Theorem 1 combines the two lemmas of this Appendix.

Appendix B
Proof of Lemma 2

The definitions of q_i and ν_i in Lemma 2 imply that the observability matrix with respect to (C_i, A_i) is of rank q_i while the null space or unobservable space is of dimension ν_i. Hence, q_i closed-loop eigenvectors span M_i, while the remaining ν_i closed-loop eigenvectors span the null space of M_i. These ν_i eigenvectors must by definition satisfy the condition of $C_i v_j^i = 0$ for all j = 1, 2, $\cdots$, ν_i. This condition implies that these eigenvectors must be eigenvectors of A_i. This can be observed by postmultiplying (16) by v_j^i for all j = 1, 2, $\cdots$, ν_i to obtain

$$A_i v_j^i=(A-DC)v_j^i+D_i C_i v_j^i. \quad \text{(B.1)}$$

This equation reduces to

$$A_i v_j^i=\lambda_j^i v_j^i+D_i C_i v_j^i \quad \text{(B.2)}$$

by (10). The last term in (B.2) is zero for all of the eigenvectors in the null space of M_i and, therefore, they must all be eigenvectors of A_i. Hence, ν_i eigenvalues and eigenvectors of A_i are specified by the solution of (13) (i.e., with respect to f_i). These eigenvalues and eigenvectors cannot be affected by the choice of D_i since the last term in (B.2) is always zero. However, the remaining q_i eigenvalues and eigenvectors are not constrained by the solution of (13) and are freely assignable by the choice of D_i. The proof of Lemma 2 is now complete. □

Appendix C
Proof of Theorem 4

This proof is due to Beard [3]. Lemma 2 implies that there are q eigenvalues that are freely assignable by D in (15) after D has been constrained to be a detection filter for all the f_i in F. Output separability implies that the eigenvalues associated with the detection space of each of the f_i can be specified independently of the other detection spaces (i.e., Lemma 3). Lemma 2 implies from (15)–(18) that ν_i eigenvalues are associated with each detection space. Condition b) of the definition of detectability requires that n eigenvalues be arbitrarily assignable when D is constrained to be a detection gain for all of the f_i in F. Hence,

$$\left(\sum_{i=1}^{r} \nu_i\right)+q=n. \quad \text{(C.1)}$$

Since $\nu \triangleq n - q$, then Theorem 4 is proven. □

Appendix D
Proof of Lemma 3

The lemma can be proven if it can be shown that the eigenvectors that span the ith detection space are independent of those that span the other detection spaces for i = 1, 2, $\cdots$, r. The independence of these sets of eigenvectors implies that there are no eigenvectors which span some overlap in the detection spaces and, hence, that the detection spaces are pairwise independent.

A proof by contradiction can be accomplished by assuming that some dependency exists among the sets of eigenvectors that span the detection spaces. For the sake of notational simplicity it is assumed that there are just two detection spaces and that the overlap between the two is a two-dimensional subspace. Furthermore, assume that v_1^1 and v_2^1 are the two eigenvectors from the ν_1-dimensional, first detection space which span the overlap. Likewise, assume that v_1^2 and v_2^2 are the two eigenvectors from the ν_2-dimensional, second detection space which also span the overlap. Since both sets of overlap eigenvectors span the same subspace, then it must be possible to write v_1^1 and v_2^1 in terms of v_1^2 and v_2^2. For instance, $v_1^1 = \omega_1 v_1^2 + \omega_2 v_2^2$ where the ω_k's are constant scalar coefficients. Premultiplication of this relationship by C (recall that $Cv_j^i \neq 0$) and the conditions of Theorem 1 produce $Cf_1 = (\omega_1 + \omega_2)Cf_2$, which is a contradiction by the assumption of output separability. These arguments are easily extended to the general case and by contradiction Lemma 3 is true. □

Appendix E
Proof of Theorem 5

The organization of this proof is as follows. First, b), c), and d) will be shown to result from (27) in a very direct and simple way. Then it will be shown that the definitions of (27) imply a), because the new excess space will not have any output component along the direction $\tilde{C}\tilde{f}_k = Cf_k$, where $\tilde{f}_k^T = [f_k^T, 0]^T$. This form of $\tilde{f}_k$ is required by the way in which a failure physically enters the problem.

The determination of the detection space of $\tilde{f}_k$ with respect to the system $(\bar{A}, \tilde{C})$ is analogous to the procedure of Section III-D. The quantities

$$\tilde{C}_k=[C_k\ \ 0],\ \tilde{K}_k=\begin{bmatrix} K_k & A_{12} \\ 0 & A_{22} \end{bmatrix} \quad \text{(E.1)}$$

are easily derived for the enlarged state space from equations of the form of (18) and (17) with $(\tilde{A}, \tilde{C})$ replacing (A, C) and where $\bar{f}_k$ replaces f_k. Now a matrix similar to (19) can be written and expanded in terms of (E.1) to obtain

$$\tilde{M}_k = \begin{bmatrix} C_k^T, (C_kK_k)^T, (C_kK_k^2)^T, \cdots, \\ 0, (C_kA_{12})^T, (C_k(K_kA_{12}+A_{12}A_{22}))^T, \cdots, \end{bmatrix}^T. \quad \text{(E.2)}$$

Because the definition of A_{12} in (27) makes A_{12} lie in the null space of M_k, the second column of terms in (E.2) is zero. Hence, the rank of $\tilde{M}_k$ is equal to that of M_k. This implies that $\tilde{\nu}_k = \tilde{n} - \tilde{q}_k = \tilde{n} - q_k = n + \nu_{ek} - q_k = \nu_k + \nu_{ek}$ and that b) is true.

A matrix similar to (23) for the enlarged system can also be defined for $\tilde{M}$. This matrix will have a form analogous to that of (E.2) and similar arguments imply that $\tilde{\nu} = \nu + \nu_{ek}$. This fact in conjunction with b) and the form of $\tilde{M}_j$ for $j \neq k$ imply that c) and, therefore, d) must be valid.

The final step is to show that a) is valid because the excess subspace has no components that lie in the kth detection space. The proof requires a detailed look at the structure of the excess subspace.

Any eigenvector, υ_e^j, which lies in the original excess subspace can be represented as a linear combination of some or all of the r failure directions plus some vector, s_j, which lies in the null space of C. Hence, $\upsilon_e^j = \Sigma_{i=1}^r \sigma_i^j f_i + s_j$ where some of the σ_i^j may be zero, and s_j cannot be zero for υ_e^j to be independent of the detection spaces eigenvectors. Furthermore, each excess eigenvector must have an s_j vector which is independent of the other s_j vectors. Therefore, V_{ek} is represented as

$$V_{ek} = \left[\sum_{I_k^1} \sigma_i^1 f_i + s_1, \cdots, \sum_{I_k^N} \sigma_i^N f_i + s_N \right] \quad \text{(E.3)}$$

where the I_k^j's for $j = 1, 2, \cdots, \nu_{ek}$ represent different subsets of the values of $i = 1, 2, \cdots, r$ where $i = k$ is by definition a member of each subset. Also, each s_j must satisfy $Cs_j = 0$ for $j = 1, 2, \cdots, \nu_{ek}$. Equation (10) for V_{ek} can be written in matrix form as $AV_{ek} = V_{ek}\Lambda_{ek} + DCV_{ek}$ which can be rewritten from (E.3) and (13) as

$$AV_{ek} = V_{ek}\Lambda_{ek} + \left[\sum_{I_k^1} \sigma_i^1 (Af_i - \bar{f}_i), \cdots, \sum_{I_k^N} \sigma_i^N (Af_i - \bar{f}_i) \right]. \quad \text{(E.4)}$$

Since the s_j for $j = 1, 2, \cdots, \nu_e$ must be independent, then another basis for the excess subspace is defined by $S_e \triangleq [s_1, \cdots, s_N]$. The subset of vectors S_{ek} in S_e which originate with V_{ek} must span the same excess subspace with respect to the new basis representation, since the same group of failure directions create that part of the excess space. The existence of a basis representation for the excess space which lies in the null space of C is guaranteed by the fact that the f_i are output separable and only r nonzero output directions can be fixed by the detection spaces. Terms from the left and right sides of (E.4) can be canceled to form

$$AS_{ek} = S_{ek}\Lambda_{ek} + \left[\sum_{I_k^1} \sigma_i^1 (\lambda_e^1 f_i - \bar{f}_i), \cdots, \sum_{I_k^N} \sigma_i^N (\lambda_e^N f_i - \bar{f}_i) \right] \quad \text{(E.5)}$$

which is now in terms of a new basis for the excess subspace due to f_k.

$\tilde{\nu}_e = \nu_e$ is now assumed. This assumption means that enlargement of the state space does not change the dimension of the excess subspace. This in turn implies that some ν_{ek}-dimensional basis, $\tilde{S}_{ek}$, must exist in the $\tilde{n}$-dimensional state space to span the excess subspace that had previously been spanned by S_{ek} in the n-dimensional state space. Such a basis can be written as $\tilde{S}_{ek} \triangleq [S_{ek}^T, I]^T$ where I is a $\nu_{ek} \times \nu_{ek}$ identity matrix which is chosen arbitrarily. If it can be shown that $\tilde{S}_{ek}$ is independent of the presence of f_k by the choice of A_{12} and A_{22} of (27), then a) will have been proven.

Equation (E.5) can be rewritten so as to be consistent with the enlarged state space dimension as

$$\tilde{A}\begin{bmatrix} S_{ek} \\ 0 \end{bmatrix} = \begin{bmatrix} S_{ek} \\ 0 \end{bmatrix}\Lambda_{ek} + \begin{bmatrix} \sum_{I_k^1} \sigma_i^1 (\lambda_e^1 f_i - \bar{f}_i), \cdots, \sum_{I_k^N} \sigma_i^N (\lambda_e^N f_i - \bar{f}_i) \\ 0 \end{bmatrix}. \quad \text{(E.6)}$$

The assumed form (Theorem 5) for $\tilde{A}$ can be expanded to obtain

$$\tilde{A}\begin{bmatrix} 0 \\ I \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix}\Lambda_{ek} + \begin{bmatrix} -\sigma_k^1 (\lambda_{ek}^1 f_k - \bar{f}_k), \cdots, -\sigma_k^N (\lambda_{ek}^N f_k - \bar{f}_k) \\ 0 \end{bmatrix} \quad \text{(E.7)}$$

and then added to (E.6) to produce

$$\tilde{A}\tilde{S}_{ek} = \tilde{S}_{ek}\Lambda_{ek} + \begin{bmatrix} \sum_{I^1} \sigma_i^1 (\lambda_e^1 f_i - \bar{f}_i), \cdots, \sum_{I^N} \sigma_i^N (\lambda_e^N f_i - \bar{f}_i) \\ 0 \end{bmatrix} \quad \text{(E.8)}$$

where the index k has now been eliminated from the I_k^j to give I^j. Because $\tilde{C}\tilde{S}_{ek} = 0$, the detection space directions of interest are given by $\tilde{A}\tilde{S}_{ek}$. Since the right-hand side of (E.8) has no component in the kth detection space, the new excess space will be independent of the presence of f_k.

The case where $\tilde{\nu}_e < \nu_e$ can be handled in exactly the same manner as the above case. The only difference here is that now one or more of the vectors in $\tilde{n} \times \nu_{ek}$-dimensional $\tilde{S}_{ek}$ will now lie in and partially span an enlarged detection space with respect to one or more f_j for $j \neq k$, rather than all of the vectors spanning the new excess space as in the previous case.

The hypothesis of a) in Theorem 5 has now been proven and the proof of Theorem 5 is complete, except for a discussion regarding the observability of $(\tilde{A}, \tilde{C})$. The enlarged system must be observable as a result of the enlargement construction of Theorem 5. This can be verified by recalling from the proof of part b) that the null space of $\tilde{M}$ must have a dimension of $\tilde{\nu} = \nu + \nu_{ek}$, and is known to be spanned by $\tilde{\nu}$ eigenvectors which either span one of the r detection subspaces or the excess subspace that exists after the enlargement process. If any one of these eigenvectors is multiplied by $\tilde{C}$, then the constraint of Theorem 1 or the nature of the excess space [e.g., see (E.3)] requires that the resulting vector be equal to one of the $\tilde{C}\tilde{f}_i$ directions for $i = 1, 2, \cdots, r$, or some linear combination thereof. Hence, the subspaces which compose the null space of $\tilde{M}$ are observable by the construction of Theorem 5, since $\tilde{C}\tilde{f}_i = Cf_i \neq 0$ for all $i = 1, 2, \cdots, r$. If $r < m$, then the remaining q eigenvectors can be freely chosen, provided that they are independent of those with respect to the r detection subspaces and the excess subspace, and are selected to span an observable subspace. Hence, the enlarged system of $(\tilde{A}, \tilde{C})$ is observable, and the proof of Theorem 5 is complete. □

Appendix F

Proof of Theorem 6

From Theorem 5 $\tilde{\nu}_k = \nu_k + \nu_{ek}$ and any increase in the dimension of some other detection space must come as a result of an equal reduction in the excess space dimension. Any increase in the dimension of ν_j to $\tilde{\nu}_j$ must come as the result of a corresponding decrease in ν_{ej} to $\tilde{\nu}_{ej}$. Hence,

$$\tilde{\nu}_j - \nu_j = \nu_{ej} - \tilde{\nu}_{ej} \tag{F.1}$$

for all $j \neq k$. Since the total increase in the dimension of those detection spaces other than that of f_k must equal the reduction in the excess space dimension, then it must be true that

$$\sum_{j=1}^{r} (\tilde{\nu}_j - \nu_j) = \sum_{j=1}^{r} (\nu_{ej} - \tilde{\nu}_{ej}) = \nu_e - \tilde{\nu}_e \tag{F.2}$$

where $j \neq k$ is a constraint on the summation terms. Equations similar to (F.1) and (F.2) can also be written for the second and successive extensions, although some of the terms in the summations of (F.2) will be zero since some directions will have been removed from the excess subspace.

For notational simplicity it is assumed here that $r = 3$. The state-space dimension must be enlarged to $\tilde{n}_r = n + \nu_{e1} + \tilde{\nu}_{e2}$ after $r - 1$ applications of Theorem 5 and where $k = 1, 2, \cdots$ on successive applications. The reason for $r - 1$ rather than r applications of Theorem 5 will become clear.

The first state-space enlargement removes the components along f_1 from the excess space and (F.2) can be employed to write $\nu_e - \tilde{\nu}_e = (\nu_{e2} - \tilde{\nu}_{e2}) + (\nu_{e3} - \tilde{\nu}_{e3})$. Substitution of this equation into the previous equation produces $\tilde{n}_r = n + (\Sigma_{k=1}^{3} \nu_{ek}) - \nu_e + (\tilde{\nu}_e - \tilde{\nu}_{e3})$ which has the form of (28) except for the final two terms. These two terms are equal and cancel each other out. Since the first enlargement removes all excess space components along f_1, the new excess space of dimension $\tilde{\nu}_e$ can have components only along f_2 and f_3. Since each excess eigenvector must have a component in two or more detection spaces (otherwise the eigenvector would lie in and span some detection space), then $\tilde{\nu}_e = \tilde{\nu}_{e2} = \tilde{\nu}_{e3}$. The second state-space enlargement will remove f_2 components from the excess space and also enlarges the f_3 detection space. Hence, the excess space will be eliminated after $r - 1$ applications of Theorem 5.

The above analysis may be repeated for an arbitrary r and, therefore, Theorem 6 is proven. □

Appendix G

Proof of Theorem 7

The conditions under which output stationarity is possible are intimately related to the relationships between the closed-loop eigenvector sets of the $\bar{v}_j^k$'s and the v_j^i's, as well as between the closed-loop eigenvalues of $\bar{\lambda}_j^k$ and λ_j^i. The $\bar{\lambda}_j^k$'s and the $\bar{v}_j^k$'s represent the closed-loop eigenvalues and eigenvectors of the filter designed as if h_k was one of the original output separable failure directions. Because of the assumption that ξ_k lies in the union of the detection spaces of the f_i for $i = 1, 2, \cdots, l$, then ξ_k and h_k can be written as in the remark after Theorem 7. The implication here is that h_k or any detection equivalent direction can be written as a unique linear combination of the eigenvectors which span the detection spaces of the f_i for $i = 1, 2, \cdots, l$.

The detection space of h_k also has a fundamental role in the development of the output stationarity conditions. The projection of the h_k detection space onto each and every detection space with respect to f_i for $i = 1, 2, \cdots$ can be shown to be of dimension $\bar{\nu}_k$ [6]. The implications of this are that there are only $\bar{\nu}_k$ eigenvalues which may be assigned to the detection space of h_k. Since the detection space of h_k has $\bar{\nu}_k$-dimensional projections on each of the l detection spaces of the f_i of (29) which comprise h_k, it may be deduced that the same arbitrarily assignable set of $\bar{\nu}_k$ eigenvalues must be a subset of the ν_i eigenvalues in each of the 1 detection spaces since $\bar{\nu}_k \leq \nu_i$. The unassignable nature of the $\nu_i - \bar{\nu}_k$ eigenvalues when $\bar{\nu}_k < \nu_i$ will become apparent later in the proof.

The temporary assumption is made here that $\bar{\nu}_k = \nu_i$ for all $i = 1, 2, \cdots, l$. The implication of equating h_k written in terms of $\bar{\alpha}_j^k$ and $\bar{v}_j^k$ from the analogous form of (11) with (30) is that

$$\bar{v}_j^k = (\bar{\delta}_j^1 v_j^1 + \cdots + \delta_j^l v_j^l)/\bar{\alpha}_j^k. \tag{G.1}$$

Hence, the structural requirements for the output stationarity of h_k are that the closed-loop eigenvectors of the detection spaces of the h_k and the f_i for $i = 1, 2, \cdots, l$ be linearly related as in (G.1). Although these arguments have been based on the assumption that $\bar{\nu}_k = \nu_i$, the proof to follow will show that the unassignable eigenvalues that occur for $\bar{\nu}_k < \nu_i$ become fixed as the result of preserving the relationship of (G.1) to fix the output direction of Ch_k.

The proof will show in the spirit of Theorem 1 that the hypothesized conditions are necessary and sufficient for the output direction of h_k to be fixed by the same detection gain as for the directions in F.

The necessity of the conditions in Theorem 7 can be shown in a manner analogous to that of th· proof of Theorem 1. An equation derived like that of (A.5) but in terms of h_k and the eigenvalues and eigenvectors with respect to its detection space is

$$\hat{\epsilon} \approx -\bar{\mu}_k \sum_{j=1}^{\bar{\nu}_k} \left(\frac{\bar{\alpha}_j^k}{\bar{\lambda}_j^k}\right)[1 - e^{\bar{\lambda}_j^k t}]C\bar{v}_j^k \tag{G.2}$$

where $C\bar{v}_j^k = Ch_k$ for a fixed output direction by Theorem 1. Similarly, an output error equation can be written for h_k in terms of the eigenvalues and eigenvectors of the detection spaces of the f_i's [i.e., from (30)] for $i = 1, 2, \cdots, l$ as

$$\hat{\epsilon} \approx -\bar{\mu}_k \left[\sum_{j=1}^{\bar{\nu}_k} \left(\frac{\bar{\alpha}_j^k}{\lambda_j^l}\right)[1 - e^{\lambda_j^l t}]C\left[\frac{1}{\bar{\alpha}_j^k}(\bar{\delta}_j^1 v_j^1 + \cdots + \delta_j^l v_j^l)\right]\right.$$
$$\left. + \sum_{i=1}^{l} \sum_{j=\nu_k+1}^{\nu_i} \left(\frac{\bar{\delta}_j^i}{\lambda_j^i}\right)[1 - e^{\lambda_j^i t}]Cv_j^i\right]. \tag{G.3}$$

The form in which (G.2) is written clearly indicates that when $\bar{\nu}_k = \nu_i$ and $\bar{\lambda}_j^k = \lambda_j^i$, then (G.1) implies that (G.3) and (G.2) are identical and the output direction of Ch_k is fixed if the directions $Cf_1, \cdots, Cf_l$ have been fixed by Theorem 1. If $\bar{\nu}_k < \nu_i$, the output stationarity of h_k can only be ensured if $\bar{\lambda}_j^k = \lambda_j^i$ for $j = 1, 2, \cdots, \bar{\nu}_k$, and the λ_j^i's of the second term in (G.3) are chosen such that the corresponding $\bar{\delta}_j^i$'s are zero. This is the reason why $\nu_i - \bar{\nu}_k$ eigenvalues must be unassignable for each detection space where $\bar{\nu}_k < \nu_i$. The necessity of the Theorem 7 conditions for output stationarity has now been shown.

The sufficiency of the conditions in Theorem 7 can also be shown in a manner similar to that of the proof of Theorem 1. An equation analogous to that of (A.1) in terms of h_k becomes $W_h = [h_k, Gh_k, \cdots, G^{n-1}h_k]$ where CW_h must be of rank 1 if the output direction Ch_k is to be fixed. An equation in terms of the h_k eigenvalues and eigenvectors can be written which is analogous to (A.2) as

$$CW_h = \left[\sum_{j=1}^{\bar{\nu}_k} \bar{\alpha}_j^k C\bar{v}_j^k, \sum_{j=1}^{\nu_k} \bar{\alpha}_j^k \bar{\lambda}_j^k C\bar{v}_j^k, \cdots, \sum_{j=1}^{\bar{\nu}_k} \bar{\alpha}_j^k \bar{\lambda}_j^{k^{\nu_k - 1}} C\bar{v}_j^k\right] \tag{G.4}$$

where the conditions of Theorem 1 would imply that CW_h is of rank 1 since $Ch_k = C\bar{v}_j^k$. Similarly, (30) can be used to rewrite W_h in terms of the eigenvalues and eigenvectors of the detection spaces of the f_i for $i = 1, 2, \cdots, l$ as

$$W_h = \left[\sum_{j=1}^{\nu_i} \sum_{i=1}^{l} \delta_j^i v_j^i, \sum_{j=1}^{\nu_i} \sum_{i=1}^{l} \delta_j^i G v_j^i, \cdots\right]. \tag{G.5}$$

If now the $\delta_j^i = 0$ for all $j > \bar{\nu}_k$ when $\nu_i > \bar{\nu}_k$, and the $\bar{\nu}_k$ freely assignable eigenvalues are chosen such that $\lambda_j^i = \bar{\lambda}_j^k$, then (G.5) can be rewritten as

$$W_h = \left[\sum_{j=1}^{\bar{\nu}_k} \sum_{i=1}^{l} \delta_j^i v_j^i, \sum_{j=1}^{\bar{\nu}_k} \sum_{i=1}^{l} \delta_j^i \bar{\lambda}_j^k v_j^i, \cdots \right] \qquad \text{(G.6)}$$

such that premultiplication by C and (G.1) produce (G.4). Hence, CW_h can be made rank 1 if the directions $Cf_1, \cdots, Cf_l$ have been fixed by Theorem 1, and the sufficiency of the Theorem 7 conditions for the output stationarity of h_k with F has been proven. □

References

[1] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, 1976.

[2] E. Y. Chow and A. S. Willsky, "Analytical redundancy and the design of robust failure detection systems," *IEEE Trans. Automat. Contr.*, vol. AC-9, pp. 603-614, July 1984.

[3] R. V. Beard, "Failure accommodation in linear systems through self-reorganization," Man-Vehicle Lab., Mass. Inst. Technol, Cambridge, MA, Rep. MVT-71-1, Feb. 1971.

[4] H. L. Jones, "Failure detection in linear systems," The Charles Stark Draper Laboratory, Cambridge, MA, Rep. T-608, Aug. 1973.

[5] B. C. Moore, "On the flexibility offered by state feedback in multivariable systems beyond closed-loop eigenvalue assignment," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 689-692, Oct. 1976.

[6] J. E. White, "Detection filter design by eigensystem assignment," Ph.D. dissertation, Univ. Texas, Austin, May 1985.

[7] V. Klema and A. Laub, "The singular value decomposition: Its computation and some applications," *IEEE Trans. Automat. Contr.*, vol. AC-25, no. 2, pp. 164-176, Apr. 1980.

[8] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*. New York: Wiley, 1971.

[9] A. J. Laub, "Numerical linear algebra aspects of control design computations," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 97-108, Feb. 1985.

[10] J. S. Meserole, Jr., "Detection filters for fault-tolerant control of turbofan engines," The Charles Stark Draper Laboratory, Cambridge, MA, Rep. T-751, June 1981.

John E. White received the B.S.A.E. degree from the University of Oklahoma, Norman, in 1976 and the M.S.A.E. and Ph.D. degrees from the University of Texas at Austin in 1980 and 1985, respectively.

From 1976 to 1979 he was employed by McDonnell Douglas Astronautics Company, St. Louis, MO. Since 1985 he has been on the Technical Staff of Sandia National Laboratories, Albuquerque, NM. His research interests include the theory and design of multiple-input/multiple-output and adaptive control systems, and failure detection and identification theory and applications.

Dr. White is a member of the American Institute of Aeronautics and Astronautics.

Jason L. Speyer (M'71-SM'82-F'85) received the B.S. degree in aeronautics and astronautics from the Massachusetts Institute of Technology, Cambridge in 1960, and the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA, in 1968.

His industrial experience includes research at Boeing, Raytheon, Analytical Mechanics Associates, and the Charles Stark Draper Laboratory. At present, he is Harry H. Power Professor in Engineering in the Department of Aerospace Engineering and Engineering Mechanics, University of Texas, Austin. He recently spent a research leave as a Lady Davis Visiting Professor at the Technion—Israel Institute of Technology, Haifa, Israel.

Dr. Speyer is presently an elected member of the Board of Governors of the IEEE Control Systems Society and Chairman of the Technical Committee on Aerospace Controls. He is a Fellow of the American Institute of Aeronautics and Astronautics.

Modeling of Parameter Variations and Asymptotic LQG Synthesis

MINJEA TAHK, MEMBER, IEEE, AND JASON L. SPEYER, FELLOW, IEEE

Abstract—**Conventional approaches in modern robustness and sensitivity theory are not adequate for the problems associated with parameter variation since the structure of parameter variations cannot be modeled properly or included in the synthesis procedure. A new modeling technique is proposed to handle a class of structured plant uncertainties in a direct way. The key is to treat deterministic parameter variations as an internal feedback loop so that the structure of parameter variations is embedded in its model. An asymptotic LQG design synthesis based on this modeling method is also presented. An important relationship between the structure of plant uncertainties and the LQG weighting matrices is obtained. This relationship clearly specifies the kind of parameter variations allowable for the LQG/LTR method.**

I. Introduction

ONE aspect of current development of MIMO (multiinput, multioutput) feedback system theory has been concerned with stability robustness and sensitivity to plant perturbations. Important developments in this field are found in the LQG/LTR (loop transfer recovery) techniques [1]-[4] and H^∞-optimization theory associated with robustness and sensitivity [5]-[8]. Although these modern techniques are useful in treating unmodeled dynamics and stochastic uncertainties such as disturbances and sensor noises, they may not be adequate in handling structured parameter variations. In their recent paper [18], Shaked and Soroka showed that an LQG controller designed by the LQG/LTR method suffers from a stability robustness problem due to a small parameter variation. Since an LQG/LTR controller is known to recover the guaranteed stability margins of an LQ regulator or a Kalman-Bucy filter [3], their result implies that a conventional usage of stability margins is no guarantee against a disastrous loss of stability.

The existing modeling methods, on which the current robustness and sensitivity studies are based, are external descriptions of the plant uncertainties, in the sense that plant uncertainties are modeled at the exterior of the plant by assigning extra blocks at the input, at the output, or around the plant as feedback or feedforward loops [11]. In practice, these modeling methods are not convenient at all in handling parameter uncertainty. Many difficulties arise from the fact that parameter uncertainties are usually given in state-space forms while the conventional uncertainty models are based on transfer function descriptions. In Section II we discuss these drawbacks in some detail and identify the inadequacy of the conventional uncertainty models for parameter uncertainty as a source of robustness problems.

Apart from the well-known modern synthesis methods, other methods have been proposed which address the robustness problem of parameter uncertainty within the framework of state-space representation. A matching condition is crucial in Lyapunov-function approaches [23], [25] so that the class of parameter variations to be considered is severely limited, although some relaxation was obtained in [24]. The matching condition assumes that a parameter variation of the state matrix A, denoted as ΔA, is spanned by the input matrix B or the output matrix C. The special characteristics of this type of parameter uncertainty will be discussed later. Another approach which deals with a larger class of parameter variation is the stochastic modeling method using state-, control-, and measurement-dependent multiplicative noises [26], [27]. This method leads to a direct synthesis which requires the coupled solution of two Riccati equations and two Lyapunov equations. However, this stochastic modeling method does not directly address the robustness problem associated with modeling errors such as parameter variations. Other synthesis methods related to parameter uncertainty are found in [28], [29]. Most of these state-space methods simply describe a parameter variation as a difference between the state-space representation of the nominal system (or, model) and that of the perturbed system (or, real system). Thereby, one objective of this paper is to better understand the role of the *structure* of parameter variations in the development of robust synthesis techniques.

This paper circumvents some of the difficulties and drawbacks of existing methods by using a modeling method which is able to characterize the structure of parameter variations in a simple way. In this method, a parameter variation is represented as an equivalent fictitious feedback loop called the *internal feedback loop* (IFL). In particular, we are using the fact that a feedback and a parameter variation are indistinguishable when input-output relations are considered. The advantages of the IFL modeling method over the existing methods are: 1) it is simple; 2) the associated stability criterion has no restriction on the closed-right-half-plane (CRHP) poles and zeros as in other methods [17]; 3) the structure of parameter variations is maintained; and 4) several modern design methods can incorporate the IFL model directly.

In IFL modeling a parameter variation ΔA is decomposed into three parts: the input, output, and feedback matrices. This decomposition is called *the input/output (I/C) decomposition*. The idea of the IFL representation or the I/O de:omposition is not new. Recently, various authors have employed this idea either implicitly or explicitly in order to study parameter uncertainty [14], [20], [30]-[32]. However, Mita and Ngamkajornvivat [33] seem to be the first to use the concept of the I/O decomposition to develop a synthesis method, which was generalized later by Shaked [34]. Their studies were limited to state-feedback problems and the major concern was pole sensitivity rather than robustness. Section II briefly discusses the IFL modeling technique which transforms a perturbed closed-system into a two-input, two-output (TITO) system. The idea of representing general plant uncertainties as a feedback loop was also previously suggested in [12], but parameter uncertainty was not explicitly treated.

The main purpose of this paper is to propose an asymptotic LQG design synthesis based on the I/O decomposition of parameter variation. Section III shows that, by selecting proper weighting matrices for the Riccati equations, either the regulator

Manuscript received March 12, 1986, revised February 11, 1987. Paper recommended by Past Associate Editor, S. P. Bhattacharyya. This work was supported in part by General Dynamics Ft. Worth Division and the Air Force Office of Scientific Research under Grant AFOSR-84-0371.

M. Tahk was with the Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, TX 78712. He is now with Integrated Systems, Santa Clara, CA 95054.

J. L. Speyer is with the Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin, Austin, TX 78712.

IEEE Log Number 8715824.

0018-9286/87/0900-0793$01.00 © 1987 IEEE

part or filter part of an LQG controller can be made asymptotically robust to a given parameter variation. In fact, there exists an explicit relationship between the LQG weighting matrices and the structure of the parameter variation. The stability robustness of the LQG control system is then determined solely by the other nonasymptotic part. This implies that an observer based on the Kalman filter can be designed by selecting suitable covariance matrices, in order to recover the robustness of the LQ regulator with respect to a given parameter variation. This asymptotical procedure generalizes the LQG LTR design to a larger class of parameter variations. This paper also shows that the class of parameter variations which can be safely treated by the LQG LTR method is limited by the structure of the input and output matrices. Some numerical results based on [18] are also given in Section IV to illustrate the advantage of this asymptotic LQG design synthesis over the conventional LQG LTR technique in the presence of parameter uncertainty.

II. Modeling of Parameter Variations

Two important classes of plant uncertainties are unmodeled dynamics (or truncated higher order dynamics) and parameter uncertainty. Since these uncertainties arise from the inaccuracy or incompleteness of mathematical models, they are often called modeling errors. In this section, we are concerned with the modeling of modeling errors rather than the modeling of real plants.

A. Drawbacks of Conventional M deling Techniques

Let $\hat{G}(s)$ and $G(s)$ be the real plant and its reduced-order model, respectively. Suppose that exact system parameters are known and the only plant uncertainty is unmodeled dynamics. For $G(s)$ to be an acceptable model, the frequency behavior of $G(s)$ should approximate that of $\hat{G}(s)$ in a reasonable manner over a certain frequency interval specified by the designer. Mathematically, the modeling error due to unmodeled dynamics can be specified in several ways. Two common models of modeling errors are

$$E_a(s) := \hat{G}(s) - G(s)$$

which is additive, and

$$E_m(s) := G(s)^{-1}[\hat{G}(s) - G(s)]$$

which is multiplicative. In practice, the exact form of $E_a(s)$ or $E_m(s)$ is neither available nor necessary. Instead, norm bounds of these error models are usually sufficient for analysis and design synthesis. Usually, unmodeled dynamics are assumed to be dominant in the high-frequency range and the norm bounds of $E_a(s)$ and $E_m(s)$ are determined in rather empirical ways.

Now consider the parameter variation case. Suppose that $\hat{G}(s)$ and $G(s)$ are of the same order but some parameter uncertainties are present, i.e., $\hat{G}(s) = G(s, \hat{p})$ and $G(s) = G(s, p)$ where p is the nominal parameter vector used in the model and $\hat{p}$ is the real parameter vector. Then, the error models $E_a(s)$ and $E_m(s)$ become

$$E_a(s) := G(s, \hat{p}) - G(s, p)$$

$$E_m(s) := G(s, p)^{-1}[G(s, \hat{p}) - G(s, p)].$$

Suppose that the parameter uncertainties are parameterized by r independent variables $\epsilon := \{\epsilon_1, \epsilon_2, \cdots, \epsilon_r\}$, i.e., $\hat{p} - p = f(\epsilon)$. This parameterization can be done easily with the state-space representations if the model and the real plant are assumed to be given as (A, B, C) and $(\hat{A}, \hat{B}, \hat{C})$, respectively. Then, the parameter uncertainties, $\Delta A = \hat{A} - A$, $\Delta B = \hat{B} - B$, and $\Delta C = \hat{C} - C$, are given as $\Delta A = \Delta A(\epsilon)$, $\Delta B = \Delta B(\epsilon)$, and $\Delta C = \Delta C(\epsilon)$. However, it is readily observed that the computation of $E_a(s, \epsilon)$ or $E_m(s, \epsilon)$ is not an easy task. Let

$$\delta_\epsilon = \|\epsilon\|, \ \delta_a = \|E_a(s, \epsilon)\|, \ \delta_m = \|E_m(s, \epsilon)\|$$

where each norm is assumed to be defined appropriately. Whil the degree of parameter variation is directly given by δ_ϵ, th relationship between δ_ϵ and δ_a, or between δ_ϵ and δ_m is extremel complicated even for a single parameter variation, except for $\delta_\epsilon = 0$ if $\delta_a = 0$ (or, $\delta_\epsilon = 0$ if $\delta_m = 0$). Apart from its complexit and inconvenience, the use of conventional error models fo parameter variations leads to a loss of information about th magnitude of parameter variation let alone the loss of its structura information.

The inadequacy of the conventional frequency-domain erro models for parameter uncertainty is important in light of stability robustness since their use in robustness analysis may lead to an incorrect conclusion on the stability robustness of a system being considered. For example, the stability margins, which are closely related to the multiplicative error model $E_m(s)$, are not useful if a small parameter variation possibly produces a very large gain variation or phase variation. For this case, a substantial amount of gain margin or phase margin cannot be a guarantee for good robustness. A good example for this situation is found in [18], as discussed in Section I.

Another drawback of the conventional methods lies in the limitations in applying stability criteria based on the conventional error models. The basic assumption of Lethomaki's stability criteria [10], [11], which is the basis of the MIMO stability margin concept, is that the perturbed plant has the same numbers of poles and zeros as the nominal plant in the CRHP. This restriction on the perturbed plant was pointed out and compared to the inverse-Nyquist-based stability criteria [17], which also assumes that the nominal plant and the perturbed plant share the same number of zeros. Thus, the class of plant uncertainties properly described by any of the conventional error models is limited by this requirement. It is important to note that those constraints on the perturbed plant result in a certain class of parameter variations, which may destabilize the system, being excluded from consideration. Therefore, we see that the unstructured plant uncertainties considered in [10] are not strictly unstructured, but there exists a definite requirement on the structure of plant uncertainty. For general parameter variations, a small parameter variation does not necessarily induce small gain and phase variations [i.e., a small $E_m(s)$], and does not necessarily keep the same number of CRHP poles and zeros.

These observations lead us to the conclusion that the conventional modeling methods for plant uncertainties and the associated stability criteria may not be a reliable tool when parameter uncertainty rather than unmodeled dynamics is involved in control system design.

B. Internal Feedback Modeling of Parameter Variations

By the I/O decomposition, a parameter variation is equivalently represented as an internal feedback loop, and then the perturbed plant is depicted as a TITO system where one feedback loop is the nominal feedback loop, and another feedback loop is for the parameter variation. This representation of parameter variation is attractive in many ways; 1) there is no restriction on the number of CRHP poles or zeros of the perturbed plant; 2) the structure of the parameter variation is easily embedded into the input matrix and output matrix of the IFL; and 3) the magnitude of the parameter variation is directly described by the magnitude of the feedback gain of the IFL.

Consider a linear, time-invariant system where

$$\dot{x} = Ax + Bu$$

$$y = Cx$$

represents a nominal system, and

$$\dot{x} = \hat{A}x + \hat{B}u$$

$$y = \hat{C}x$$

represents its perturbed system. The vectors $x \in \Re^n$, $u \in \Re^m$, and $y \in \Re^l$ denote the state, the input, and the output, respectively. We assume that $\Delta A := \hat{A} - A \neq 0$, but $\Delta B := \hat{B} - B = 0$ and $\Delta C := \hat{C} - C = 0$, i.e., only the state matrix is subject to variation. It will be shown that ΔB or ΔC can be embedded by an approximation procedure into a ΔA of an augmented system.

Suppose that ΔA is parameterized as a function of r variables $\epsilon = \{\epsilon_1, \cdots, \epsilon_r\}$ and given as

$$\Delta A(\epsilon) = \sum_{i=1}^{r} S_i \epsilon_i$$

where the S_i's are constant matrices. By decomposing S_i as $S_i = M_i N_i$, we can rewrite ΔA as

$$\Delta A(\epsilon) = -ML(\epsilon)N$$

where $M \in \Re^{n \times p}$ and $N \in \Re^{q \times n}$ are constant matrices determined by M_i's and N_i's, and $L(\epsilon) \in \Re^{p \times q}$ is a matrix function of $\epsilon = \{\epsilon_1, \cdots, \epsilon_r\}$. The decomposition described above is called here an I/O decomposition of ΔA. Note that the I/O decomposition is not unique. To avoid problems with an unnecessarily large M or N, we assume that the decomposition is such that M and N are full rank; i.e., M and N are of minimal dimensions. This nonuniqueness is not important in stability analysis and design synthesis, as shall be discussed later.

Given an I/O decomposition of $\Delta A(\epsilon)$, the perturbed plant can be written as

$$\dot{x} = Ax + Bu + Mw$$

$$y = Cx$$

$$z = Nx$$

$$w = -L(\epsilon)z$$

where two variables z and w are introduced, respectively, as an auxiliary output and an input connected to the internal feedback loop with a gain $L(\epsilon)$, as shown in Fig. 1(a) where

$$\bar{g} := \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \begin{bmatrix} C\phi B & C\phi M \\ N\phi B & N\phi M \end{bmatrix}$$

and $\phi = (sI - A)^{-1}$.

Let $K(s)$ be a compensator. Then, it is easy to see that the perturbed closed-loop system shown in Fig. 1(b) is equivalent to Fig. 1(c) where

$$\begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix}$$

and

$$\bar{G} := \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$$

$$= \begin{bmatrix} (I + g_{11}K)^{-1} g_{11} K & (I + g_{11}K)^{-1} g_{12} \\ g_{21}(I + Kg_{11})^{-1} K & g_{22} - g_{21} K (I + g_{11}K)^{-1} g_{12} \end{bmatrix}.$$

For a given ϵ, a sufficient condition for closed-loop stability is that

$$\det [I + \alpha L(\epsilon) G_{22}(j\omega)] \neq 0$$

for all $\alpha \in [0, 1]$ and $\omega \in \Re$ [12], [14]. When only parameter uncertainty is considered, the above criteria gives not only input-output stability but also the internal stability of the closed-loop system [19]. In other words, the stability of the unobservable or uncontrollable modes, which cannot be studied by the conventional methods, are also examined. This stability criterion is also independent of the I/O decomposition. Suppose that $-ML(\epsilon)N$ is an I/O decomposition of the ΔA. Then

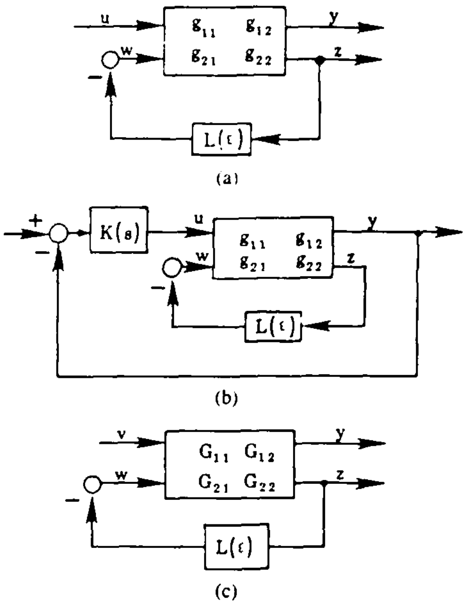


Fig. 1. TITO system representations of the perturbed system.

$$\det [I + \alpha L(\epsilon) G_{22}(j\omega)]$$

$$= \det [I + \alpha L(\epsilon) N(\phi - \phi BK(I + g_{11}K)^{-1} C\phi) M]$$

$$= \det [I + \alpha M L(\epsilon) N(\phi - \phi BK(I + g_{11}K)^{-1} C\phi)]$$

$$= \det [I - \alpha \Delta A(\phi - \phi BK(I + g_{11}K)^{-1} C\phi)].$$

Therefore, the lack of uniqueness of the decomposition does not affect this stability criterion.

C. Parameter Variations of the Matrices B and C

For parameter variations in B and C, the I/O decompositions can be obtained in several ways. One way is to construct the state-space representation of the closed-loop system (A_c, B_c, C_c) and $(\hat{A}_c, \hat{B}_c, \hat{C}_c)$, and to obtain an I/O decomposition for $\Delta A_c = \hat{A}_c - A_c$. However, this method requires a state-space representation of the feedback compensators. Or we can construct augmented state-space representations (A_a, B_a, C_a) and $(\hat{A}_a, \hat{B}_a, \hat{C}_a)$ where $\hat{B}_a = B_a$, $\hat{C}_a = C_a$ but $\Delta A_a = \hat{A}_a - A_a$ approximates ΔB and ΔC. The latter is more favorable since a unified approach is possible for all kinds of parameter variations so it will be discussed in detail in this section. Some other methods for ΔB and ΔC are also found in [20], [36].

Suppose that the ith column of B is subject to a perturbation. Then we augment the state vector x by an extra state x_b which follows the input u_i fast enough so that its dynamics are negligible when compared to other modes of A. Effectively, u_i becomes a state, and the plant is approximated by

$$\begin{bmatrix} \dot{x} \\ \dot{x}_b \end{bmatrix} = \begin{bmatrix} A & b_i \\ 0 & -\sigma \end{bmatrix} \begin{bmatrix} x \\ x_b \end{bmatrix} + \begin{bmatrix} B_o \\ b_o \end{bmatrix} u$$

where σ is a sufficiently large positive number, b_i is the ith column of the matrix B, and

$$B_o = B - [0\ 0 \cdots b_i \cdots 0]$$

$$b_0 = [0\ 0 \cdots \sigma \cdots 0].$$

In fact, the transfer function from u_i to x_b becomes

$$x_b(s) = \frac{\sigma}{s+\sigma} u_i(s), \qquad \sigma \gg 1.$$

Similarly, for the jth row of C, which is subject to a perturbation, we can augment the state space as

$$\begin{bmatrix} \dot{x} \\ \dot{x}_c \end{bmatrix} = \begin{bmatrix} A & 0 \\ c_j & -\tau \end{bmatrix} \begin{bmatrix} x \\ x_c \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u$$

$$y = [C_o \; c_o] \begin{bmatrix} x \\ x_c \end{bmatrix}$$

where $\tau \gg 1$, c_j is the jth row of C, and

$$C_o = C - [0\; 0 \cdots c_j^T \cdots 0]^T$$

$$c_o = [0\; 0 \cdots \tau \cdots 0]^T.$$

This augmentation gives

$$y_i(s) = \frac{\tau}{s+\tau} x_c(s), \qquad \tau \gg 1.$$

It is easy to prove that controllability and observability are not affected by the above state augmentation procedures. The poles and zeros of the original system are also those of the augmented system, and the only alteration is the addition of some poles on the negative real axis, which are well beyond the bandwidth of the plant. The proofs are simple and left to the reader. Since ΔB and ΔC can be approximated as ΔA_a of an augmented system, it is assumed from now on that the perturbed plant does not have perturbations in B and C, i.e., $\hat{B} = B$ and $\hat{C} = C$.

III. An Asymptotic LQG Design Synthesis

This section introduces an asymptotic LQG design procedure based on the internal feedback modeling method described in the previous section. It is shown that the finite poles of either the regulator or filter of an LQG control system can be designed to be asymptotically insensitive to a specified parameter variation via a suitable selection of LQG weighting matrices. These weighting matrices turn out to be closely related to the structure of a particular class of parameter variations. The robustness of the LQG control system is essentially determined by the remaining sensitive part of the LQG compensator. In other words, the robustness of the LQG control system recovers either the robustness of the LQ regulator or that of a Kalman–Bucy filter. The robustness problem again reduces to determining the robustness of either the LQ regulator or Kalman–Bucy filter. Furthermore, the LQG LTR procedure may be considered a special case of this asymptotic LQG method, although the explicit connection between plant uncertainties and the LQG weighting matrices had not been made.

A. Similar Parameter Variations

[illegible] structured parameter variations has great [illegible] of the asymptotic behavior of LQG poles. [illegible] characteristics are given here.

[illegible] that $M_1 \in \Re^{n\times m_1}$, $M_2 \in \Re^{n\times m_2}$, $N_1 \in$ [illegible] Then M_1 is said to be *column-similar* to M_2 [illegible] and N_1 is said to be *row-similar* [illegible] N_2 [illegible] exists a matrix P such that $M_1 =$ [illegible] M_2 and that there exists a matrix [illegible] similar to N_2.

Definition 2 (Similar Parameter Variations): Consider a nominal plant (A, B, C) and its perturbed plant $(\hat{A}, B, C)$ where $\Delta A = \hat{A} - A = -MLN$. Then, ΔA is said to be *input-similar* if M is column-similar to B, and it is said to be *output-similar* if N is row-similar to C. It is also called *similar* if it is input-similar or output-similar.

An important characteristic of similar variations is that they do not perturb the zeros of the plant. Let Z_t and Z_d be the sets of transmission zeros and decoupling zeros of the nominal plant (A, B, C), respectively. Similarly, we define $\hat{Z}_t$ and $\hat{Z}_d$ as the sets of transmission zeros and decoupling zeros of the perturbed plant $(\hat{A}, B, C)$. (We will follow the definition of zeros given in [16], [22].)

Lemma 1: If ΔA is similar (i.e., input-similar or output-similar), then

$$Z_t \cup Z_d = \hat{Z}_t \cup \hat{Z}_d.$$

Proof: For input-similar variations,

$$\text{rank} \begin{bmatrix} sI - A + BPLN & B \\ -C & 0 \end{bmatrix} = \text{rank} \begin{bmatrix} sI - A & B \\ -C & 0 \end{bmatrix}.$$

Therefore, the system zeros are invariant. The proof is similar for output-similar variations. □

B. Asymptotic Pole Sensitivities

Consider an output feedback system with a constant gain feedback given as

$$\dot{x} = Ax + Bu$$

$$y = Cx$$

$$u = v - kFy$$

where $B \in \Re^{n\times m}$, $C \in \Re^{m\times n}$, $F \in \Re^{m\times m}$ are assumed to be full rank, and $k \in \Re$. The numbers of the input and output are assumed to be the same.

Lemma 2: Suppose that $G(s) = C(sI - A)^{-1}B$ has j_t finite transmission zeros. Then j_t poles of the closed-loop system, $A - kBFC$, asymptotically approach the transmission j_t zeros of $G(s)$ as $k \to \infty$.

This is a well-known feedback property for single-input, single-output (SISO) systems. Lemma 2 implies that the number of the finite eigenvalues of $A - kBFC$ as $k \to \infty$ is equal to $j_t + j_d$ where j_d is the number of decoupling zeros corresponding to unobservable and/or uncontrollable modes and j_t is the number of asymptotically finite closed-loop poles. For the MIMO case, its formal proof can be found in [21, Theorem 4.3].

Theorem 1: Consider the above output feedback system. Suppose that the state matrix A is perturbed by a parameter variation given as $\Delta A = -MLN$. Then, as $k \to \infty$, every finite eigenvalue of $A - kBFC$ is asymptotically insensitive to the parameter variation $-MLN$ if ΔA is similar.

Proof: Consider a similar variation $-\beta MLN$ where $\beta \in [0, 1]$. Let $P(\beta, k)$ be the set of finite eigenvalues of $A - \beta MLN - kBFC$ and $Z(\beta)$ be the set of zeros of $(A - \beta MLN, B, C)$. For the nominal system ($\beta = 0$), we denote them as $P(0, k)$ and $Z(0)$, respectively. Lemma 2 implies that $P(\beta, k) \to Z(\beta)$ and $P(0, k) \to Z(0)$ as $k \to \infty$. However, fron Lemma 1, $Z(\beta) = Z(0)$. Thus, $\lim_{k\to\infty} P(\beta, k) = \lim_{k\to\infty} P(0, k)$. Suppose that $p \in P(0, k)$ and $\hat{p} \in P(\beta, k)$ approach $z \in Z(0)$ as $k \to \infty$. Then, there exists a constant $k(\delta)$ for any $\delta > 0$ such that $|z - p| < \delta$ and $|z - \hat{p}| < \delta$ for all $k > k(\delta)$. Therefore, $|p - \hat{p}| < 2\delta$ for $k > k(\delta)$. The continuity of eigenvalues of $A - \beta MLN - kBFC$ with respect to β completes the proof. □

C. Asymptotic Robustness of a Full State Regulator with Observer

Theorem 1 states that the finite poles are insensitive to a parameter variation when ΔA is similar. This property is not readily useful since the closed-loop stability is not guaranteed by simple output feedback. However, the results of Theorem 1 give some insight into the asymptotic robustness of observer-based control systems. Consider a full-state feedback regulator with observer (FSRO) for which the state matrix of the closed loop system is given by

$$\hat{A}_c = \begin{bmatrix} \hat{A} & BK_c \\ -K_f C & A - BK_c - K_f C \end{bmatrix}$$

where $\hat{A} = A - MLN$, $A \in \mathfrak{R}^{n\times n}$, $B \in \mathfrak{R}^{n\times m}$, $C \in \mathfrak{R}^{l\times n}$, $M \in \mathfrak{R}^{n\times p}$, $N \in \mathfrak{R}^{q\times n}$, $L \in \mathfrak{R}^{p\times q}$, and $K_c \in \mathfrak{R}^{m\times n}$ and $K_f \in \mathfrak{R}^{n\times l}$ are the regulator and observer gain, respectively. For the rest of this paper, we assume that B, C, K_f, K_c, M, and N are of full rank and (A, B, C) is minimal. If $\Delta A = 0$, then the closed-loop poles are determined by $A - BK_c$ and $A - K_f C$. The choice of K_c only affects the regulator part while K_f affects only the observer part. If $\Delta A \neq 0$ (i.e., $\hat{A} \neq A$), then the perturbed closed-loop poles are no longer determined by $A - BK_c$ and $A - K_f C$. However, the following theorem shows that the coupling between the regulator and the filter can be eliminated asymptotically.

Theorem 2: Consider the above FSRO system with a parameter variation $\Delta A = -MLN$. Suppose that

$$K_f = \gamma \tilde{K}_f + K_{fo}(\gamma)$$

where $\tilde{K}_f$ is a finite matrix and $(1/\gamma)K_{fo}(\gamma) \to 0$ as $\gamma \to \infty$. Suppose that M is column-similar to $\tilde{K}_f$. Then, as $\gamma \to \infty$: 1) $\hat{A} - BK_c$ determines half of the closed-loop poles; 2) there are asymptotic poles approaching the zeros of $(A, \tilde{K}_f, C)$; and 3) these asymptotic poles are asymptotically insensitive to the parameter variation $\Delta A = -MLN$.

Proof: Since

$$\hat{A}_c = \begin{bmatrix} \hat{A} & BK_c \\ 0 & A - BK_c \end{bmatrix} - \begin{bmatrix} 0 \\ \tilde{K}_f + \frac{1}{\gamma} K_{fo}(\gamma) \end{bmatrix} \gamma [C \;\; C],$$

Lemma 2 implies that, as $\gamma \to \infty$, the finite closed-loop poles approach the zeros of

$$H(s) := [C \;\; C] \begin{bmatrix} sI - \hat{A} & -BK_c \\ 0 & sI - A + BK_c \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \tilde{K}_f + \frac{1}{\gamma} K_{fo}(\gamma) \end{bmatrix}.$$

Let

$$A_0 = \begin{bmatrix} \hat{A} & BK_c \\ 0 & A - BK_c \end{bmatrix}, \quad C_0 := [C \;\; C],$$

$$B_0 := \begin{bmatrix} 0 \\ \tilde{K}_f + \frac{1}{\gamma} K_{fo}(\gamma) \end{bmatrix}, \quad B_1 := \begin{bmatrix} 0 \\ \tilde{K}_f \end{bmatrix}, \quad B_2 := \begin{bmatrix} 0 \\ \frac{1}{\gamma} K_{fo}(\gamma) \end{bmatrix}.$$

Define

$$P_0(s) := \begin{bmatrix} sI - A_0 & B_0 \\ -C_0 & 0 \end{bmatrix}, \quad P_1(s) := \begin{bmatrix} sI - A_0 & B_1 \\ -C_0 & 0 \end{bmatrix},$$

$$P_2(s) := \begin{bmatrix} 0 & B_2 \\ 0 & 0 \end{bmatrix}.$$

Then, $P_0(s) = P_1(s) + P_2(s)$. Suppose that z_o is a finite zero of $P_0(s)$. Let $\underline{\sigma}(\cdot)$ and $\bar{\sigma}(\cdot)$ denote the minimum singular value and maximum singular value, respectively. Then, $\sigma(P_0(z_o)) = 0$. From the singular-value inequality $\underline{\sigma}(A) - \bar{\sigma}(B) \leq \underline{\sigma}(A + B) \leq \underline{\sigma}(A) + \bar{\sigma}(B)$, we have

$$\underline{\sigma}(P_0(z_o)) - \bar{\sigma}(P_2(z_o)) \leq \underline{\sigma}(P_1(z_o)) \leq \underline{\sigma}(P_0(z_o)) + \bar{\sigma}(P_2(z_o)).$$

Then, $\underline{\sigma}(P_1(z_o)) \to 0$ as $\gamma \to \infty$ since $\lim_{\gamma\to\infty} \bar{\sigma}(P_2(z_0)) = 0$. Therefore, z_0 is also a zero of $P_1(s)$ in the limit so that the finite closed-loop poles are determined by the zeros of $P_1(s)$. Let

$$H_1(s) := C_0(sI - A_0)^{-1}B_1.$$

Then, a simple calculation gives

$$H_1(s) = C(sI - \hat{A})^{-1}\{I + MLN(sI - A + BK_c)^{-1}\}\tilde{K}_f.$$

Since M is column-similar to $\tilde{K}_f$, there exists a matrix P such that $M = \tilde{K}_f P$. Then, $H_1(s)$ can be written as $H_1(s) = H_f(s)H_c(s)$ where

$$H_f(s) := C(sI - \hat{A})^{-1}\tilde{K}_f$$

$$H_c(s) := I + PLN(sI - A + BK_c)^{-1}\tilde{K}_f.$$

Consider $H_c(s)$ first. Since

$$\begin{aligned} \det[H_c(s)] &= \det[I + PLN(sI - A + BK_c)^{-1}\tilde{K}_f] \\ &= \det[I + (sI - A + BK_c)^{-1}\tilde{K}_f PLN] \\ &= \det(sI - A + BK_c)^{-1} \det(sI - A + BK_c + \tilde{K}_f PLN) \\ &= \det(sI - A + BK_c)^{-1} \det(sI - \hat{A} + BK_c), \end{aligned}$$

the zeros of $H_c(s)$ are the eigenvalues of $\hat{A} - BK_c$. This shows that all the nominal regulator poles, which are determined by $A - BK_c$, are perturbed to the eigenvalues of $\hat{A} - BK_c = A - MLN - BK_c$. This proves part 1). For the observer part, Lemma 1 implies that the zeros of $H_f(s) = C(sI - \hat{A})^{-1}\tilde{K}_f$ are the zeros of $C(sI - A)^{-1}\tilde{K}_f$ since M is input-similar for the system $(A, \tilde{K}_f, C)$. Thus, the asymptotically finite poles of the observer part of the perturbed system go to the zeros of $C(sI - A)^{-1}\tilde{K}_f$, which completes the proof of part 2). Finally, using the same arguments as used in Theorem 1, we can easily show that part 3) is true.□

The following properties are direct from Theorem 2: 1) the asymptotic pole locations are determined by the dominant part of the observer gain, $\tilde{K}_f$; and 2) the regulator poles are perturbed in the same way as a state-feedback system $A - BK_c$ is perturbed by ΔA when ΔA is column-similar to $\tilde{K}_f$. Since the asymptotic observer poles are insensitive to ΔA in this case, we may suspect that the robustness of the FSRO controller is determined by its regulator part. This is true, but we have not considered the sensitivities of asymptotic infinite poles yet. This robustness property becomes explicit in Theorem 3 to follow. We first consider the next lemma for its proof.

Lemma 3: Suppose that A, B, C are matrices such that $A + B + C$ is invertible. Then,

$$A + B(A + B + C)^{-1}C = (A + C)(A + B + C)^{-1}(A + B).$$

Proof: Let $D = (A + B + C)^{-1}$, then

$$\begin{aligned} A + BDC &= ADD^{-1} + BDC \\ &= ADA + ADB + ADC + BDC. \end{aligned}$$

Using $(A + B + C)DC = C$ and $CD(A + B + C) = C$, we obtain

$$ADC + BDC = CDA + CDB.$$

Therefore,

$$\begin{aligned} ADA+ADB+ADC+BDC &= ADA+ADB+CDA+CDB \\ &= (A+C)D(A+B). \qquad \square \end{aligned}$$

Theorem 3: Consider the robustness function $G_{22}(s)$ of the asymptotic FSRO control system of Theorem 2. Suppose that the nominal closed-loop is stable and $C(sI - A)^{-1}\tilde{K}_f$ has no zeros in the CRHP. Then, as $\gamma \to \infty$,

$$G_{22}(j\omega) \to N(j\omega I - A + BK_c)^{-1}M$$

pointwise for all $\omega \in \Re$.

Proof: For an FSRO compensator, $K(s)$ is given as

$$K(s) = K_c(sI - A + BK_c + K_fC)^{-1}K_f.$$

Thus, the robustness function $G_{22}(s)$ becomes

$$\begin{aligned} G_{22}(s) &= N\phi M - N\phi BK(I+GK)^{-1}C\phi M \\ &= N(sI-A+BK_c(sI-A+BK_c+K_fC)^{-1}K_fC)^{-1}M \\ &= N(sI-A+BK_c)^{-1}(sI-A+BK_c+K_fC) \\ &\quad \cdot (sI-A+K_fC)^{-1}M \\ &= N(sI-A+BK_c)^{-1}M + N(sI-A+BK_c)^{-1}BK_c \\ &\quad \cdot (sI-A+K_fC)^{-1}M \end{aligned}$$

where Lemma 3 is used in the third equality, and all the inverses exist along the imaginary axis since the closed-loop is stable. Now $M = \tilde{K}_fP$ since M is column-similar to $\tilde{K}_f$. Let $\tilde{K}_f = (1/\gamma)K_f$. From $K_f = \gamma\tilde{K}_f + K_{fo}(\gamma)$, we obtain $M = \hat{K}_fP - (1/\gamma)K_{fo}(\gamma)P$. Then,

$$\begin{aligned} &K_c(sI-A+K_fC)^{-1}M \\ &= K_c[\phi - \phi\gamma\hat{K}_f(I+\gamma C\phi\hat{K}_f)^{-1}C\phi]\left(\hat{K}_fP - \frac{1}{\gamma}K_{fo}(\gamma)P\right) \\ &= K_c\phi\hat{K}_f[I-\gamma(I+\gamma C\phi\hat{K}_f)^{-1}C\phi\hat{K}_f]P \\ &\quad - K_c[\phi - \phi\gamma\hat{K}_f(I+\gamma C\phi\hat{K}_f)^{-1}C\phi]\frac{1}{\gamma}K_{fo}(\gamma)P \\ &= K_c\phi\hat{K}_f(I+\gamma C\phi\hat{K}_f)^{-1}P - \frac{1}{\gamma}K_c\phi K_{fo}(\gamma)P \\ &\quad + K_c\phi\hat{K}_f(I+\gamma C\phi\hat{K}_f)^{-1}C\phi K_{fo}(\gamma)P \end{aligned}$$

where $\phi = (sI - A)^{-1}$. For the time being, suppose that A has no eigenvalues on the imaginary axis. Since $C\phi\tilde{K}_f$ has no CRHP zeros, $\underline{\sigma}(C\phi\tilde{K}_f) > 0$ along the imaginary axis. Using the singular value inequality used in the proof of Theorem 2, we can show that

$$\lim_{\gamma\to\infty} \underline{\sigma}(C\phi\hat{K}_f) = \underline{\sigma}(C\phi\tilde{K}_f).$$

Again,

$$\underline{\sigma}(\gamma C\phi\hat{K}_f) - 1 \le \underline{\sigma}(I+\gamma C\phi\hat{K}_f) \le \underline{\sigma}(\gamma C\phi\hat{K}_f) + 1.$$

Thus, $\underline{\sigma}(I + \gamma C\phi\hat{K}_f) \to \infty$ as $\gamma \to \infty$ since $\underline{\sigma}(\gamma C\phi\hat{K}_f) = \gamma\underline{\sigma}(C\phi\hat{K}_f) \to \infty$ as $\gamma \to \infty$. Then, as $\gamma \to \infty$,

$$\bar{\sigma}(I+\gamma C\phi\hat{K}_f)^{-1} = \underline{\sigma}^{-1}(I+\gamma C\phi\hat{K}_f) \to 0.$$

Also,

$$\begin{aligned} \bar{\sigma}[(I+\gamma C\phi\hat{K}_f)^{-1}C\phi K_{fo}(\gamma)] &\le \bar{\sigma}(I+\gamma C\phi\hat{K}_f)^{-1}\bar{\sigma}(C\phi K_{fo}(\gamma)) \\ &\sim \underline{\sigma}^{-1}(C\phi\hat{K}_f)\bar{\sigma}\left(C\phi\frac{1}{\gamma}K_{fo}(\gamma)\right) \to 0 \end{aligned}$$

as $\gamma \to \infty$. Therefore, we see that

$$K_c(j\omega I - A + K_fC)^{-1}M \to 0 \qquad \text{as } \gamma \to \infty$$

and

$$\lim_{\gamma\to\infty} G_{22}(j\omega) = N(j\omega I - A + BK_c)^{-1}M$$

for all $\omega \in \Re$. Finally, we need to consider the case for which A has some eigenvalues on the imaginary axis so that $\phi = (sI - A)^{-1}$ is not defined for $\omega \in \Re$. However, this difficulty can be avoided as follows. Let $A' = A + K_oC$ be a matrix without eigenvalues on the imaginary axis. Since (A, C) is assumed to be observable, there always exist such a K_o. Then,

$$sI - A + K_fC = sI - A' + (K_f + K_o)C = sI - A' + K_f'C$$

where $K_f' = \gamma\tilde{K}_f + K_{fo}(\gamma) + K_o = \gamma\tilde{K}_f + K_{fo}'(\gamma)$. Then, we can apply the same procedure as above. □

The robustness function $G_{22}(j\omega)$ given in Section II-C determines the robustness of the FSRO control system. However, $N(sI - A + BK_c)^{-1}M$ is indeed the robustness function of a state-feedback regulator subject to ΔA. (It can be shown easily by constructing a TITO system for the state-feedback regulator.) Therefore, Theorem 3 implies that a regulator with an asymptotic full-order estimator recovers the robustness of a regulator with full-state feedback. Furthermore, we see that the sensitivities of asymptotic infinite poles of the observer part do not contribute to the robustness. Although they may be sensitive to ΔA, these infinite poles cannot be perturbed to the CRHP by ΔA. An exact dual of the above asymptotic property exists as the regulator gain K_c instead of K_f becomes infinitely large (i.e., $K_c \equiv \beta\tilde{K}_c$ where $\beta \to \infty$, and $\tilde{K}_c$ is chosen as $N = Q\tilde{K}_c$ where Q is a finite matrix).

If $p > l$ and M is full rank, we cannot find a $\tilde{K}_f$ such that $M = \tilde{K}_fP$. Similarly, if $q > m$ and N is full rank, there does not exist $\tilde{K}_c$ such that $N = Q\tilde{K}_c$. Therefore, the above asymptotic property cannot be obtained if $p > l$ and $q > m$, i.e., the rank of the input matrix B and the output matrix C, which are assumed to be full rank, limit the rank of M and N in applying the asymptotic, partial desensitization procedure described above.

In Theorem 3, the conditions seem to be redundant since we only use the fact that $A - BK_c$ and $A - K_fC$ do not have pure imaginary eigenvalues and $C(sI - A)^{-1}\tilde{K}_f$ is not singular along the imaginary axis. However, the closed-loop system should be stable for any robustness issue to be meaningful. Also, $C(sI - A)^{-1}\tilde{K}_f$ should not have CRHP zeros since those zeros determine the asymptotic closed-loop poles. In the next section, the asymptotic observer with closed-loop stability can be obtained via a Kalman-Bucy filter (KBF) design.

D. Asymptotic LQG Design Synthesis

Consider a KBF problem stated as

$$\dot{x} = Ax + Bu + R\xi$$

$$y = Cx + \nu\eta$$

$$\dot{\hat{x}} = A\hat{x} + Bu + K_f(y - C\hat{x})$$

with a cost

$$J = \lim_{T\to\infty} \frac{1}{T} E\left[\int_0^T (x-\hat{x})^T(x-\hat{x})\,dt\right]$$

where ξ and η are unit-strength Gaussian noises of appropriate dimensions. Under the assumption that (A, R) is stabilizable and (A, C) is detectable, the optimal filter gain K_f is obtained as

$$K_f = \frac{1}{\upsilon^2}\Sigma C^T$$

where K_f is the positive definite solution to a Riccati equation

$$\Sigma A^T + A\Sigma^T + RR^T - \frac{1}{v^2}\Sigma C^T C\Sigma = 0.$$

Theorem 4: If there exists a full rank $R \in \Re^{n \times l}$ such that M is column-similar to R, (A, R) is stabilizable, and $C(sI - A)^{-1}R$ has no CRHP zero, then, as $v \to 0$, the asymptotic properties of Theorem 3 are obtained.

Proof: As $v \to 0$, $K_f \to (1/v)RW$ where W is a nonsingular matrix, if (A, R) is stabilizable, (A, C) is detectable, and $C(sI - A)^{-1}R$ is minimum phase. The proofs of these properties are given in [3], [9], [15] and not repeated here. Let $\gamma = (1/v)$. Then we see that the asymptotic gain assumes the form K_f as $K_f = \gamma RW + K_{fo}$ where $(1/\gamma)K_{fo} \to 0$ as $\gamma \to \infty$. Let $\tilde{K}_f = RW$. Then, if M is column-similar to R, there exists a matrix P such that $M = RP$ and $M = RP = \tilde{K}_f W^{-1} P$. Therefore, M is also column-similar to $\tilde{K}_f$. Finally, the closed-loop stability is automatically satisfied by a LQG design. □

Suppose that $s = z_o$ is a zero of (A, M, C). Then the system matrix

$$P(z_o) = \begin{bmatrix} z_o I - A & M \\ -C & 0 \end{bmatrix}$$

is rank deficient, and there exist vectors $p_1 \in \Re^{n \times 1}$ and $p_2 \in \Re^{p \times 1}$ such that

$$(z_o I - A)p_1 + Mp_2 = 0, \qquad Cp_1 = 0.$$

Since span $\{M\} \subset$ span $\{R\}$, we can find a vector $p_3 \in \Re^{l \times 1}$ such that

$$(z_o I - A)p_1 + Rp_3 = 0, \qquad Cp_1 = 0.$$

This implies that the zeros of (A, M, C) are also the zeros of (A, R, C). In other words, $C(sI - A)^{-1}R$ is minimum-phase only if $C(sI - A)^{-1}M$ is minimum-phase. Therefore, we can apply the asymptotic LQG synthesis only for minimum-phase $C(sI - A)^{-1}M$.

The dual of Theorem 4 is obtained by considering a LQR problem but is not treated here. The importance of Theorem 4 is that there exists a direct structural relationship between the parameter variation and the optimal LQG weighting matrices for robustness (M and R for the above asymptotic procedure). As far as an observer design is concerned, these results can be interpreted as follows: 1) given M, which gives partial information on the structure of ΔA, we can choose $\gamma^2 RR^T$ as the covariance of the process noise (a natural choice is $R = M$ if $p = 1$); and 2) as $\gamma \to \infty$, the stability robustness is determined through the regulator gain K_c (Theorem 3) while the nominal observer poles become insensitive to ΔA. Although the robustness of the LQ part depends on the regulator gain K_c and ΔA, a general theory is not available for finite regulator gains. However, it is possible to apply the asymptotic procedures to the regulator and observer part at the same time. Using a similar method to the one used in Theorem 3, we can prove that absolute robustness (i.e., $G_{22} = 0$) is obtained asymptotically if M and N are used for the observer and regulator design, respectively. Finally, we can show that the nonuniqueness of the I/O decomposition is irrelevant to the asymptotic LQG design. For example, if MLN is an I/O decomposition of ΔA, so is $MT_1T_1^{-1}LT_2^{-1}T_2N$. Then, we see that an R satisfying the conditions of Theorem 4 for M also satisfies them for MT_1.

E. Remarks on the LQG/LTR Method

This section shows that LQR/LTR is a special case of the asymptotic weighting strategy discussed in the previous section. Two types of LQG/LTR procedures, which are dual to each other, have been studied in various works [1]–[4]. One is *sensitivity recovery* where the loop transfer function of the KBF is recovered at the output by an asymptotic regulation of the output $y = Cx$. The other is *robustness recovery* where the loop transfer function of LQR is recovered at input by an observer design based on an asymptotic KBF with a white noise injected at the input (i.e., $R = B$). Note that observer insensitivity to parameter variations requires that M be column-similar to R. Therefore, in the LQG/LTR method, the asymptotic finite poles are guaranteed to be insensitive only to either input-similar or output-similar parameter variations. For example, the robustness recovery procedure makes the observer part insensitive to input-similar variations, and the stability robustness to such variations is solely determined by the regulator part. Thus, the LQG/LTR procedure works well if the structure of parameter variation is related to the structure of the input matrix B or the output matrix C, i.e., ΔA is similar for the (A, B, C). However, if $\Delta A = -MLN$ is neither input-similar nor output-similar, all the closed-loop poles are perturbed when the LQG/LTR procedure is used, and there is no guaranteed stability robustness to the parameter variation in spite of the guaranteed stability margins. It is noted that the matching condition of the Lyapunov-function methods [23], [25] requires that parameter variations be similar. This observation may imply that robust stabilization with similar parameter variations is less complicated than with other types of parameter variations.

The relationship between the error models of Section II-A and similar parameter variations is now discussed. For an input-similar variation ($M = BP$)

$$\begin{aligned}\hat{g}_{11} &= g_{11} - g_{12}L(I + g_{22}L)^{-1}g_{21} \\ &= C\phi B - C\phi BPL(I + N\phi BPL)^{-1}N\phi B \\ &= g_{11}(I + PLN\phi B)^{-1}.\end{aligned}$$

The perturbed plant is expressed as a multiplication of the plant and a transfer function placed at the input, which is similar to the conventional uncertainty model $E_m(s)$ discussed in Section II-A. Similarly, an output-similar parameter variation has a structural similarity to a conventional uncertainty model placed at the output. It is noted that these multiplicative forms cannot be obtained if ΔA is not similar. We consider the robustness function G_{22} associated with the input-similar parameter variations to examine the implications of this structural similarity. Suppose that M is input-similar. Then there exists a matrix P such that $M = BP$, and

$$\begin{aligned}G_{22} &= N\phi M - N\phi BK(I + C\phi BK)^{-1}C\phi M \\ &= N\phi BP - N\phi BK(I + C\phi BK)^{-1}C\phi BP \\ &= N\phi B(I + KG)^{-1}P\end{aligned}$$

where $G = C\phi B$. Since P and $N\phi B$ are arbitrary for general input-similar variations, the robustness to input-similar variations is optimized by minimizing $\bar{\sigma}[(I + KG)^{-1}]$, or equivalently maximizing $\underline{\sigma}[(I + KG)]$. Also, the robustness to output-similar parameter variations can be improved by increasing $\underline{\sigma}[(I + KG)]$. These singular values are, in fact, the robustness measures used in the conventional singular-value methods [10], [11]. From this observation, we can conclude that the robustness to the uncertainty modeled at the input (at the output, respectively) is equivalent to the robustness of the class of input-similar (output-similar, respectively) variations. The close relationship between similar parameter variations and conventional multiplicative uncertainties also shows the limitation of the existing multiplicative uncertainty models in modeling parameter variations that only similar variations are properly represented by these models. Therefore, the LQG/LTR methods based on these uncertainty models may fail when applied to a general parameter variation problem.

IV. Numerical Examples

We consider the example presented in [18]. The plant is given by

$$A=\begin{bmatrix}-1 & 0 & 0\\ 0 & -2 & 0\\ 0 & 0 & -3\end{bmatrix}\quad B=\begin{bmatrix}1\\ -2\\ 1\end{bmatrix}$$

$$C=[3\ 3\ 3]\quad \Delta C=[\epsilon\ 0\ 0].$$

The transfer functions of the nominal system and perturbed system are given as

$$G(s)=\frac{6}{(s+1)(s+2)(s+3)},\quad \hat{G}(s)=\frac{6+\epsilon(s+2)(s+3)}{(s+1)(s+2)(s+3)},$$

respectively. It is noted that an arbitrarily small parameter variation creates a pair of zeros at infinity. For $\epsilon > 0$, these zeros are located at $s = \pm\infty$ along the real axis. If $\epsilon < 0$, then we have two complex zeros at $s = -2.5 \pm j\infty$. Although these zeros affect very little the low-frequency characteristics of the plant, the plant becomes very uncertain in the high frequency range since they introduce large gain and phase variations.

By using the state augmentation procedure described in Section II-C we obtain the augmented system such that

$$\hat{A}_a=\begin{bmatrix}-1 & 0 & 0 & 0\\ 0 & -2 & 0 & 0\\ 0 & 0 & -3 & 0\\ 3+\epsilon & 3 & 3 & -\tau\end{bmatrix}\quad B_a=\begin{bmatrix}1\\ -2\\ 1\\ 0\end{bmatrix}\quad C_a^T=\begin{bmatrix}0\\ 0\\ 0\\ \tau\end{bmatrix}$$

$$M=\begin{bmatrix}0\\ 0\\ 0\\ 1\end{bmatrix}\quad N^T=\begin{bmatrix}-1\\ 0\\ 0\\ 0\end{bmatrix}$$

where $\tau = 1000$ for computation. For the LQ part, the cost is given by

$$J=\lim_{T\to\infty}\frac{1}{T}\int_0^T (z^Tz+\rho^2u^Tu)\,dt$$

where the weightings are chosen as $z = y = Cx$ and $\rho = 0.1$. This produces three nominal regulator poles $s = -4.52$ and $s = -2.26 \pm j2.87$. For the filter part, the weighting matrix is given as

$$Q_f=w_{fm}^2MM^T+w_{fb}^2B_aB_a^T.$$

If $w_{fm} = 0$ and $w_{fb} \to \infty$, then we have a robustness recovery procedure, which is one of the LQG/LTR procedures. The weighting $w_{fb} = 0$ and $w_{fm} \to \infty$ corresponds to the asymptotic LQG based on the structure of the parameter variation.

Figs. 2–4 show how the sensitivity of the LQG system is reduced by using the structure of the parameter variation in selecting the weighting matrix of the filter Riccati equation. The far-left poles induced by the state augmentation are not shown in these figures. In Fig. 2 the root loci are plotted for the robustness recovery procedure when $|\epsilon| \leq 0.5$. The filter poles are shown to be very sensitive as pointed out in [18]. The sensitivity of the LQG/LTR procedure is due to the sensitivity of the filter because M is not similar to B_a. Fig. 3 shows that these poles become less sensitive when w_{fm} increases from 0.0 to 10.0. The sensitivity is considerably reduced by the addition of the weighting $w_{fm}^2MM^T$ associated with the parameter variation ΔA_a. Fig. 4 also shows that the filter poles are completely insensitive when M is used instead of B_a for the weighting matrix. In this example, the three filter poles $s = -1, -2, -3$ are the open-loop poles. These poles correspond to stabilizable but uncontrollable modes of the pair (A_a, M) of the augmented system. They are the decoupling zeros of $C_a(sI - A_a)^{-1}M$, and, by Theorem 2, are insensitive to ΔA_a as $w_{fm} \to \infty$.

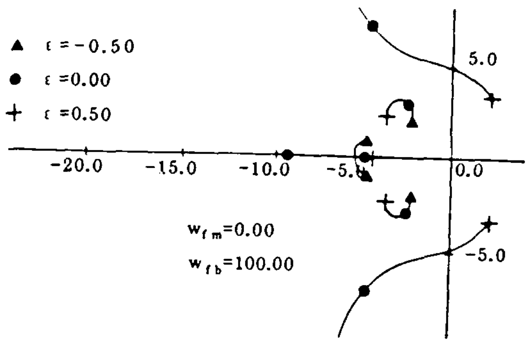


Fig. 2. Root loci of the robustness recovery procedure.

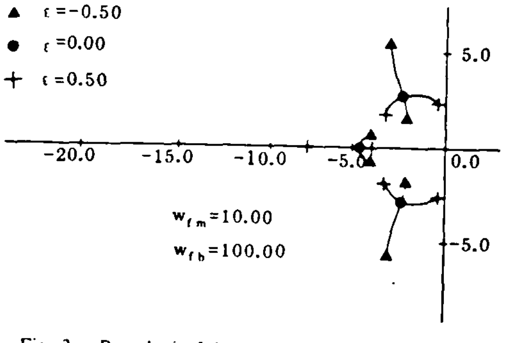


Fig. 3. Root loci of the LQG design using B and M.

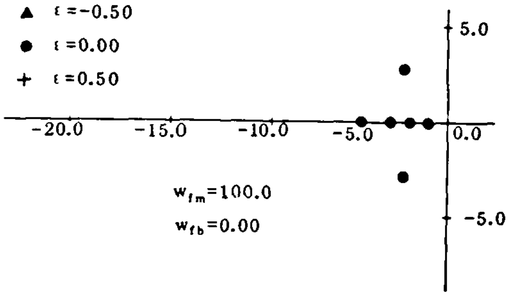


Fig. 4. Root loci of the asymptotic LQG design using M.

Note that the regulator poles are also insensitive in Fig. 4. By a direct evaluation of $K(s)$ we can show that $K(s) \to 0$ as $w_{fm} \to 0$. This implies that the optimal robustness for this example is obtained without feedback rather than using high feedback gain. It is because the output matrix C of the original problem is assumed to be totally uncertain (i.e., only M is used for the filter-part weighting) while any variation of C does not perturb the poles of the plant. This result is valid, if trivial, as far as robustness is concerned. The matrix N can be used for the asymptotic regulator without resulting in a trivial $K(s)$ as shown in Fig. 5. In this case, the use of N corresponds to the assumption that the first column of the augmented A matrix is uncertain. The regulator state weighting matrix now used is

$$Q_c=w_{cn}^2N^TN+w_{cc}^2C_a^TC_a.$$

If $w_{cc} \to \infty$ and $w_{cn} = 0$, then the LQG/LTR robustness property is obtained at the system output. If $w_{cc} = 0$ and $w_{cn} \to \infty$, then the regulator is made insensitive to the parameter error in C.

V. Conclusions

In this paper, the plant uncertainties are divided into two groups: unstructured uncertainties such as unmodeled dynamics

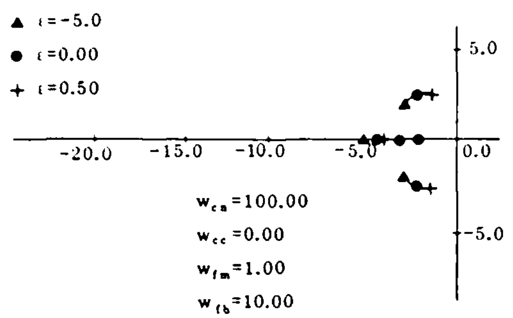


Fig. 5. Root loci of the asymptotic LQG design using N.

and structured uncertainties such as parameter variations. These two groups are compared to each other, and the inadequacy of the current modeling methods for parameter variations is discussed. Based on the I O decomposition of parameter variations, a new modeling method is then introduced to solve this problem. An asymptotic LQG procedure is proposed for desensitizing the closed-loop poles when parameter variations are present. The class of parameter variations to which an asymptotic LQG is robust (or insensitive) is determined by the structure of the weighting matrices. This implies that, if the structure of a parameter variation is known, then the robustness can be improved by this asymptotic LQG procedure. As shown in Section IV the robustness of the overall system is not affected by the filter part but determined by the controller part, which is independent of the asymptotic filter gain K_f. Therefore, for parameter variations, the LQG synthesis for robustness reduces to a synthesis problem of the nonasymptotic LQ controller (or the Kalman–Bucy filter in the dual problem). It is then shown that the LQG/LTR procedure is a special case of this asymptotic approach. In fact, the guaranteed stability margins of LQG/LTR controllers are not meaningful for the general class of parameter variations considered here but for the even smaller class of variations such that the parameter structures are similar to the input matrix or output matrix of the plant.

References

[1] J. C. Doyle and G. Stein, "Robustness with observers," *IEEE Trans. Automat. Contr.*, vol. AC-24, Aug. 1979.

[2] G. Stein, "Generalized quadratic weights for asymptotic regulator properties," *IEEE Trans. Automat. Contr.*, vol. AC-24, Aug. 1979.

[3] J. C. Doyle and G. Stein, "Multivariable feedback design: Concepts for a classical/modern synthesis," *IEEE Trans. Automat. Contr.*, vol. AC-26, Feb. 1981.

[4] G. Stein and M. Athans, "The LQG/LTR procedures for multivariable feedback control design," Mass. Inst. Technol., Cambridge, MA, Rep. 1984.

[5] G. Zames, "Feedback and optimal sensitivity; Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Trans. Automat. Contr.*, vol. AC-26, Apr. 1981.

[6] G. Zames and B. A. Francis, "Feedback, minimax sensitivity and optimal robustness," *IEEE Trans. Automat. Contr.*, vol. AC-28, May 1983.

[7] G. Zames and D. Bensoussan, "Multivariable feedback, sensitivity and decentralized control," *IEEE Trans. Automat. Contr.*, vol. AC-28, Nov. 1983.

[8] B. A. Francis and G. Zames, "Design of H_∞-optimal multivariable feedback systems," in *Proc. 22nd CDC*, Dec. 1983.

[9] C. A. Harvey and G. Stein, "Quadratic weights for asymptotic regulator properties," *IEEE Trans. Automat. Contr.*, vol. AC-23, June 1978.

[10] N. A. Lethomaki, N. R. Sandell, Jr., and M. Athans, "Robustness results in linear-quadratic Gaussian based multivariable control designs," *IEEE Trans. Automat. Contr.*, vol. AC-26, Feb. 1981.

[11] N. A. Lethomaki, D. A. Castanon, B. C. Levy, G. Stein, N. R. Sandell, Jr., and M. Athans, "Robustness and modeling error characterization," *IEEE Trans. Automat. Contr.*, vol. AC-29, Mar. 1984

[12] J. C. Doyle, "Analysis of feedback systems with structured uncertainties," *IEE Proc.*, Nov. 1982

[13] J. C. Doyle, "Synthesis of robust controllers and filters," in *Proc. 22nd CDC*, Dec. 1983.

[14] J. C. Willems, "Almost invariant subspaces. An approach to high gain feedback designs—Part I. Almost controlled invariant subspaces," *IEEE Trans. Automat. Contr.*, vol. AC-26, Feb. 1981.

[15] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. New York: Wiley, 1972.

[16] H. H. Rosenbrock, *Computer-Aided Control System Design*. New York: Academic, 1974.

[17] I. Postlewaite, M. S. Tombs, Y. K. Foo, and A. P. Loh, "On the relationship between Lethomaki's robustness test and an inverse Nyquist based test," *IEEE Trans. Automat. Contr.*, vol. AC-30, Sept. 1985.

[18] U. Shaked and E. Soroka, "On the stability of the continuous-time LQG optimal control," *IEEE Trans. Automat. Contr.*, vol. AC-30, Oct. 1985.

[19] M. Tahk and J. L. Speyer, "On the robustness of multivariable systems with structured plant uncertainties," Univ. Texas at Austin, Rep., Mar. 1985.

[20] B. G. Morton and R. M. McAfoos, "A mu-test for robustness analysis of a real-parameter variation problem," in *Proc. ACC*, June 1985.

[21] B. Kouvaritakis and U. Shaked, "Asymptotic behavior of root-loci of linear multivariable systems," *Int. J. Contr.*, vol. 23, no. 3, pp. 297–340, 1976.

[22] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[23] G. Leitmann, "Guaranteed asymptotic stability for some linear systems with bounded uncertainties," *J. Dynam. Syst., Measurement Contr.*, vol. 101, pp. 212–216, 1979.

[24] B. R. Barmish and G. Leitman, "On ultimate boundedness control of uncertain systems in the absence of matching conditions," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 253–258, 1982.

[25] A. R. Galimidi and B. R. Barmish, "The constrained Lyapunov problem and its application to robust output feedback stabilization," *IEEE Trans. Automat. Contr.*, vol. AC-31, May 1986.

[26] D. E. Gustafson and J. L. Speyer, "Linear minimum variance filters applied to carrier tracking," *IEEE Trans. Automat. Contr.*, vol. AC-21, 1976.

[27] D. S. Bernstein and D. C. Hyland, "The optimal projection equations for reduced-order modeling, estimation and control of linear systems with stratonovich multiplicative white noises," *SIAM J. Contr.*, submitted for publication.

[28] J. R. Sesak, P. Linkins, and T. Coradetti, "Flexible spacecraft control by modal error sensitivity suppression," *J. Astronautical Sci.*, pp. 131–156, Apr.-June 1979.

[29] A. Ashkenazy and A. E. Bryson, Jr., "Control logic for parameter insensitivity disturbance attenuation," *J. Guidance Contr.*, pp. 383–388, July-Aug. 1982.

[30] P. Molander, "Stabilization of uncertain systems," Dept. Automat. Contr., Lund Instit. Technol., Rep. LUTFD2/(TRFT-1020)/1-111/(1979), Aug. 1979.

[31] J. L. Willems and J. C. Willems, "Robust stabilization of uncertain systems," *SIAM J. Contr. Opt.*, pp. 352–375, 1983.

[32] I. R. Peterson, "A Riccati equation approach to the design of stabilizing controllers and observers for a class of uncertain systems," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 904–907, Sept. 1985.

[33] T. Mita and K. Ngamkajornvivat, "On the design of a system having zero-sensitivity poles," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 601–602, Aug. 1976.

[34] U. Shaked, "The design of multivariable system having zero sensitive poles," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 117–119, Feb. 1979.

[35] J. Doyle, R. E. Wall, and G. Stein, "Performance and robustness analysis for structured uncertainty," in *Proc. 21st CDC*, 1982.

[36] B. R. Barmish, "Stabilization of uncertain systems via linear control," *IEEE Trans. Automat. Contr.*, vol. AC-28, Aug. 1983.

Minjea Tahk (S'85–M'85–M'86) was born in Taegu, Korea, on March 17, 1954. He received the B.S. degree from Seoul National University, Seoul, Korea, in 1976 and the M.S. and Ph.D. degrees from the University of Texas at Austin in 1983 and 1986, all in aerospace engineering.

During the period 1976 to 1981 he was a Research Engineer at the Agency for Defense Development, Daejeon, Korea. From 1981 to 1986 he was a Research Assistant with the Guidance Control Group in the Department of Aerospace Engineering, University of Texas at Austin. He is currently with Integrated Systems, Santa Clara, CA. His research interests include the areas of robust control, flexible structure control, and design of flight control systems.

Jason L. Speyer (M'71–SM'82–F'85), for a photograph and biography, see p. 603 of the July 1987 issue of this TRANSACTIONS.

END
DATE
FILMED
DTIC
4/88